A Two-Stage Shared Component Model for Modelling Multiple Correlated Exposures and Their Health Effects

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the School of Public Health
University of Saskatchewan
Saskatoon

By
Xi Chen

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Director of School of Public Health
Health Sciences Building E-Wing, 104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 2Z4
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

In many environmental and epidemiological studies, the observed exposures are often highly correlated. Estimating the exposure-specific effects of the multiple correlated exposures on the health outcome in a statistical model can be challenging since multicollinearity among the exposures can lead to biased estimators. This study proposed a two-stage shared component model for addressing this challenge to utilize the information of the collected exposures fully. The first stage is a pollution model in which the shared and residual components are obtained to represent the common and unique effects from each correlated explanatory variable. The second stage is a disease model that the shared and residual components were included as explanatory variables for modelling the disease risk. The proposed model is motivated by an environmental health study that investigated the association between air pollutants and the respiratory hospital admissions in Greater Glasgow, Scotland, in 2011. The three highly correlated pollutants $PM_{2.5}$, PM_{10} and NO_2 are simultaneously modelled in the two-stage shared component model. Our results indicated that the air pollutants jointly increased the respiratory disease risk while NO_2 has a stronger health effect. We also investigated the finite sample properties of the proposed two-stage shared component model, which demonstrated that the proposed method could help resolve the issue of multicollinearity with appropriate and easily interpretable coefficients.

Acknowledgements

Foremost I would like to express my sincere gratitude to my supervisor, Dr. Cindy Feng, for her academic and financial support as well as her encouragement and patience throughout the period of my MSc study. Without her guidance and advice, this thesis could not have been accomplished. It is my great honour to work under her supervision.

I would like to thank Dr. Punam Pahwa and Dr. June Lim on my advisory committee for their insightful comments and encouragement. I would also like to appreciate Professor Dr. Ducan Lee from the University of Glasgow for allowing me to use the pollution data from his research.

I am grateful to the School of Public Health for the academic and financial support, and many thanks to all the professors, graduate students, and staff in the School of Public Health.

Finally, I would especially like to express my very profound gratitude to my family for providing me with continuous love and support during my whole life. I would also like to thank my friends who give me a lot of help and support during my daily life in Saskatoon these years.

Thanks for everything.

Contents

Permission to Use i

Abstract ii

Acknowledgements iii

Contents iv

List of Tables vi

List of Figures vii

List of Abbreviations ix

1 Introduction 1
- 1.1 Overview . . . 1
- 1.2 Motivating Example . . . 2
- 1.3 Research Objectives and Questions for the Study . . . 3
- 1.4 Outline of the Thesis . . . 3

2 Literature Review 4
- 2.1 Multicollinearity . . . 4
- 2.2 Impact of Multicollinearity . . . 4
- 2.3 Diagnosis of Multicollinearity . . . 7
- 2.4 Conventional Strategies for Solving Multicollinearity . . . 8
  - 2.4.1 Variable Selection . . . 9
  - 2.4.2 Principal Component Analysis . . . 11

3 Methodology 13
- 3.1 Model Specification . . . 13
  - 3.1.1 Stage 1—Pollution Model . . . 13
  - 3.1.2 Stage 2—Disease Model . . . 14
- 3.2 Model Validation and Evaluation . . . 15
  - 3.2.1 Root Mean Squared Error . . . 15
  - 3.2.2 Akaike's Information Criterion . . . 15
  - 3.2.3 Residual Plots . . . 16

4 Application: Glasgow Study 17
- 4.1 Motivating Study . . . 17
- 4.2 Exploratory Data Analysis . . . 19
- 4.3 Model Formulation . . . 20

4.3.1 Stage 1 - Pollution Model . . . 20
4.3.2 Stage 2 - Disease Model . . . 21
4.4 Results . . . 22
4.4.1 Stage 1 - Pollution Model . . . 22
4.4.2 Stage 2 - Disease Model . . . 26

5 Simulation Study **37**
5.1 Data Generation . . . 37
5.2 Results . . . 39

6 Discussion and Future Work **47**
6.1 Discussion . . . 47
6.2 Limitations and Future Work . . . 48

Bibliography **49**

Appendix A Results of the Two-stage Shared Component Model with Different Exposure Variable Entered the Stage 1 Model First **55**
A.1 PM_{10} . . . 55
A.2 NO_2 . . . 55

Appendix B Simulation Results with Negatively Correlated Explanatory Variables **57**

Appendix C Research Ethics Approval **63**

List of Tables

4.1 The estimated regression coefficients and variance components of the shared and residual components of the stage 1 model. 23
4.2 A summary of the estimated parameters and the overall fit of the Poisson models. 30
4.3 A summary of the estimated parameters and the overall fit of the negative binomial models. 30

5.1 The average of the estimated correlation coefficient (r) and variance inflation factor (VIF) for the explanatory variables from the 1000 simulated samples. 39

A.1 The estimated parameters in exponential term and the overall fit of the proposed two-stage model 55
A.2 The estimated parameters in exponential term and the overall fit of the proposed two-stage model 56

B.1 The average of the estimated correlation coefficient (r) and variance inflation factor (VIF) for the explanatory variables from the 1000 simulated samples. 57

List of Figures

4.1 The distribution of the observed number of hospital admissions for respiratory diseases over the 271 administrative units in Greater Glasgow, Scotland, in the year 2011. . . . 18
4.2 Map of the standardized incidence ratio (SIR) of hospital admissions for respiratory diseases over the 271 administrative units in Greater Glasgow, Scotland, in the year 2011. . . . 18
4.3 The histogram of the yearly concentrations of the three pollutants over the 271 administrative units in Greater Glasgow, Scotland, in the year 2010. . . . 19
4.4 The correlations between scaled $PM_{2.5}$, PM_{10} and NO_2 after square root transformation. . . . 20
4.5 The residuals versus fitted values for x_2 (left plot) and x_3 (right plot) based on the State 1-pollution model. . . . 23
4.6 Maps of the normalized yearly concentrations of pollutants. . . . 25
4.7 Maps of the shared and residual components of the Stage 1 model. . . . 25
4.8 The scree plot from PCA. . . . 27
4.9 The observed disease cases vs. fitted disease cases from the Poisson models. 31
4.10 The observed disease cases vs. fitted disease cases from the negative binomial models. . . . 32
4.11 Plots of RQR vs. fitted number of disease count from the Poisson models. . 33
4.12 Plots of RQR vs. fitted number of disease count from the negative binomial models. . . . 34
4.13 The Q-Q plots of RQR for the Poisson models . . . 35
4.14 The Q-Q plots of RQR for the negative binomial models. . . . 36

5.1 The average of the estimated regression coefficients of the proposed two-stage shared component model (top panel), the naive model (middle panel) and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. 42
5.2 The average of the standard errors of the regression coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. . . . 43
5.3 The probability of statistically significant coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. . . . 44
5.4 The averaged RMSE of the predicted response variable based on the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples. . . . 45
5.5 The averaged AIC of the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples. . 46

B.1 The average of the estimated negative regression coefficients of the proposed two-stage shared component model (top panel), the naive model (middle panel) and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. 58
B.2 The average of the standard errors of the regression coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. 59
B.3 The probability of statistically significant coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples. 60
B.4 The averaged RMSE of the predicted response variable based on the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples. 61
B.5 The averaged AIC of the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples. . 62

List of Abbreviations

PCA	Principal Components Analysis
VIF	Variance Inflation Factor
RMSE	Root Mean Squared Error
AIC	Akaike's Information Criterion
JSA	Job Seekers Allowance
AQI	Air Quality Index
RQR	Randomized Quantile Residuals

1. Introduction

1.1 Overview

Air pollution has been a global public health issue, as a growing body of epidemiological evidence has shown that air pollutants are responsible for numerous diseases, such as respiratory and cardiovascular diseases [1, 2, 3, 4]. The commonly measured pollutants include volatile organic compounds, total particulate matter, nitrogen oxides, sulphur oxides, carbon monoxide and ozone [1]. A study in Pittsburgh investigating the lethal impact of a few different air pollutants found a consistent positive association between PM_{10} and daily mortality among the age group under 75 and a significant association between dew point and daily mortality for the order-age group [5]. A few other cohort studies revealed that a higher risk of lung cancer was relating to air pollution exposures, especially sulphate and particulates [1, 2]. Besides, other cancers were also found associated with air pollution; for instance, increased concentration of NO_2 may raise the risk of postmenopausal breast cancer [6]. Moreover, adverse health events like impaired cardiac or cardiovascular function [3, 4], reduced lung function [2], preterm delivery [7] and low birth weight [8] were also proved to be associated with ambient air pollution.

Although numerous studies have found evidence of the relationship between air pollution and diseases, many of these previous studies only considered one or two pollutants in their analyses [3, 6, 8, 9]. Discarding information of other air pollutants is not ideal, since air pollution is a complex mixture of several different pollutants with varying concentrations and composition in each region [10]. Although modelling the impact of multiple air pollutants simultaneously on the health outcomes is desirable, simply including all the correlated risk factors will raise the problem of multicollinearity. Multicollinearity refers to the situation that an independent variable is highly correlated with the other independent variable(s) in the regression equation. It has no effect on overall model fit and the predictions of the model, but can cause unstable standard errors and p-values of estimated regression coefficients for

the correlated covariates [11, 12]. Thus, when conclusions are made from studies with highly correlated risk factors in the same model, erroneous and misleading conclusions may be drawn due to the presence of multicollinearity among the covariates.

Traditional approaches to address the issue due to multicollinearity include variable selection, principal component analysis (PCA), ridge regression [11, 13], etc. Bayesian kernel machine regression [14] is proposed in recent years to examine the overall effect of multiple correlated pollutants. Composite air quality index created by a combination of multiple pollutants after transformed to a common scale [15] is also used in environmental health studies to circumvent the multicollinearity issue. However, each of these approaches has its own limitations. Specifically, the procedure of variable selection may drop the risk factor(s) that we wish to study. PCA is not favourable on epidemiological interpretation [11]. Ridge regression leads to biased estimates of the coefficients [13]. Composite air quality index and Bayesian kernel machine regression in environmental studies focus on the joint effect rather than the marginal effect [14, 15], so the health effect of an individual pollutant is unknown.

1.2 Motivating Example

This thesis is motivated by an environmental health study for modelling the impact of air pollutants on respiratory diseases in the city of Glasgow in 2011 [16]. To circumvent the issue of multicollinearity, the concentrations of a few highly correlated pollutants were separately modelled to predict the counts of hospital admissions. Then $PM_{2.5}$, PM_{10} and NO_2 were found related to the increased risks of respiratory diseases individually in the studied area [16].

The goal of our study is to simultaneously model the impact of the highly correlated air pollutants on the disease incidence. To this end, we proposed a two-stage shared component model to address the problem of multicollinearity. Not only the overall effect of air pollution, but also the marginal effect of each pollutant can be studied in our proposed model. The first stage is a pollution model, which aims to find a series of independent variables representing the multiple pollutants. The second stage is a disease model, which can estimate the health effects of the multiple pollutants via the variables from the first stage.

1.3 Research Objectives and Questions for the Study

The research objectives are to

1. Develop a modelling method to simultaneously study the respective effect of highly correlated air pollutants on the disease incidence;

2. Conduct simulation study to evaluate the properties of the proposed modelling method in comparison to the traditional methods (the naive model including all the pollutants and the PCA method) for overcoming multicollinearity issue.

The research questions are

1. Does the proposed model generate more stable parameter estimates as compared to the naive model including all the pollutants?

2. Does the proposed model offer reasonable epidemiological interpretation on the effect of multiple air pollutants in contrast to the PCA method?

1.4 Outline of the Thesis

In the following chapters of this thesis, Chapter 2 reviews the issue of multicollinearity and some of the existing solutions; Chapter 3 introduces our proposed methodology; Chapter 4 presents the results of the analysis on the real motivating data by comparing the proposed method with the traditional methods for resolving the issue of multicollinearity, and Chapter 5 conducts a simulation study to investigate the properties of the proposed methods; Chapter 6 presents the summary and future work.

2. Literature Review

In view of our purpose to simultaneously estimate the effect of multiple highly correlated explanatory variables free of multicollinearity, the definition of multicollinearity and literatures overcoming multicollinearity are reviewed.

2.1 Multicollinearity

Multiple regression model is often developed in data processing of health studies, to analyse the health effects of various factors relating to human life. Usually, the interested health response is expected to link with multiple explanatory variables, among which some may also be correlated to each other besides the response. Multicollinearity refers to a situation in which two or more independent variables in a multiple regression model are highly linearly related [11].

2.2 Impact of Multicollinearity

The correlation between explanatory variables could inflate the standard errors of regression coefficients, thus lead to insignificant coefficients [11], and sometimes it may change the coefficients to opposite sign of their true values [17]. Previous research showed that the higher the correlation between explanatory variables, the further the estimates and the standard errors of the regression coefficients depart from their true values [12]. The unreliable and biased regression coefficients caused by multicollinearity make the interpretation of the covariate effects unrealistic and unconvinced [12]. However, when the focus of the study is to predict the response variable, multicollinearity among the exposures is not of concern, as previous studies showed that multicollinearity does not affect the overall model fit [18].

For example, in the context of a multiple linear regression model, let Y_i denotes the response variable, $i = 1, \cdots, n$, where n denotes the sample size and $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iK})$ denotes the vector of the covariates for the i^{th} subject, and ϵ_i is the unobserved random error

following a normal distribution with mean zero and variance σ^2. The linear regression model is expressed as,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \epsilon_i \,, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)$ denotes the regression coefficients for the explanatory variables.

The least-squares estimators $\hat{\boldsymbol{\beta}}$ of the regression coefficients obtained from Equation 2.1 are:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{2.2}$$

Suppose the explanatory variables are standardized and partitioned as $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \cdots\ \boldsymbol{X}_K]$, with $\boldsymbol{X'X}$ being the $K \times K$ symmetric correlation matrix,

$$\boldsymbol{X'X} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1K} \\ & 1 & r_{23} & \cdots & r_{2K} \\ & & & & \vdots \\ & & & & r_{K-1,K} \\ & & & & 1 \end{bmatrix}.$$

where r_{ij} is the correlation coefficient between the i^{th} and the j^{th} explanatory variables.

Then consider a $K \times K$ symmetric matrix $\boldsymbol{A}$, which is diagonalizable that there exists an orthogonal matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_K]$ to make it to a diagonal matrix $\boldsymbol{D}$.

$$\boldsymbol{U'AU} = \boldsymbol{D} = diag(\lambda_1, \lambda_2, \cdots, \lambda_K). \tag{2.3}$$

By definition, $\boldsymbol{U}$ is the matrix of eigenvectors of $\boldsymbol{A}$, and the elements on the diagonal of $\boldsymbol{D}$, namely λs, are the eigenvalues of $\boldsymbol{A}$. We take $\boldsymbol{A} = \boldsymbol{X'X}$, therefore,

$$\boldsymbol{D} = \boldsymbol{U'X'XU} = diag(\lambda_1, \lambda_2, \cdots, \lambda_K)\,. \tag{2.4}$$

Then we have

$$\boldsymbol{X'X} = \boldsymbol{U} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ & 0 & & \lambda_K \end{bmatrix} \boldsymbol{U'}.$$

As the matrix $\boldsymbol{U}$ is orthogonal, its inverse is equal to its transpose. When we take the inverse of both sides,

$$(\boldsymbol{X'X})^{-1} = \boldsymbol{U} \begin{bmatrix} 1/\lambda_1 & & & 0 \\ & 1/\lambda_2 & & \\ & & \ddots & \\ 0 & & & 1/\lambda_K \end{bmatrix} \boldsymbol{U'}.$$

The trace of $(\boldsymbol{X'X})^{-1}$ is equal to the sum of variances of the regression coefficients apart from the error variance σ^2. For $\boldsymbol{U}$ is an orthogonal matrix that $\boldsymbol{UU'}$=$\boldsymbol{U'U}$=$\boldsymbol{I}$, and from matrix algebra $\texttt{tr}(\boldsymbol{U'AU}) = \texttt{tr}\ \boldsymbol{UU'A}$. We can infer that,

$$\texttt{tr}(\boldsymbol{X'X})^{-1} = \texttt{tr}\ \boldsymbol{U'U} \begin{bmatrix} 1/\lambda_1 & & & 0 \\ & 1/\lambda_2 & & \\ & & \ddots & \\ 0 & & & 1/\lambda_K \end{bmatrix}$$

$$= \texttt{tr} \begin{bmatrix} 1/\lambda_1 & & & 0 \\ & 1/\lambda_2 & & \\ & & \ddots & \\ 0 & & & 1/\lambda_K \end{bmatrix}$$

$$= \sum_{k=1}^{k} \frac{1}{\lambda_k}.$$

When no collinearity exists between the explanatory variables, all the eigenvalues of the correlation matrix will be 1. In contrast, at least one eigenvalue is near zero when there is multicollinearity. We then consider the effect of multicollinearity in terms of the coefficient variance:

$$\frac{E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} = \texttt{tr}(\boldsymbol{X'X})^{-1} = \sum_{k=1}^{k} \frac{1}{\lambda_k}. \tag{2.5}$$

From Equation (2.5) we can see that when any of the eigenvalues λ_k are small that nearly equal to 0, $E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be large.

The effect of multicollinearity on the coefficient estimates can also be inferred from Equation (2.5):

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) - (\boldsymbol{\beta}'\boldsymbol{\beta}),$$

so

$$E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) = (\boldsymbol{\beta}'\boldsymbol{\beta}) + \sigma^2 \sum_{k=1}^{k} \frac{1}{\lambda_k}. \tag{2.6}$$

Equation (2.6) shows that if λ_k is small due to multicollinearity, the vector of estimated coefficients $\hat{\boldsymbol{\beta}}$ will be heavily biased from the true values $\boldsymbol{\beta}$ [19].

2.3 Diagnosis of Multicollinearity

Variance inflation factor (VIF) is commonly used to measure the severity of multicollinearity, which indicates how much the variance of the explanatory variable is inflated due to multicollinearity in explaining the variance of the response variable. VIF can be computed as,

$$VIF_k = \frac{1}{1 - R_k^2}, \tag{2.7}$$

where R_k^2 denotes the squared correlation between the k^{th} covariate and other covariates, namely, the proportion of variance that the k^{th} covariate shared with other covariates in the regression model [20]. The rule of thumb to decide severe multicollinearity is ad hoc. Generally, VIF greater than 10 indicates that the multicollinearity can not be ignored, but some other researchers suggest that a small VIF can also raise concerns for the validity of the analysis, such as inflated type I error [15] for coefficients estimation. The threshold of VIF equal to 5 or 8 is also adopted in some studies [11, 12].

Although multicollinearity can lead to biased inferences on the parameter estimates in a regression model, a review by Vatcheva et al. found that only a very small number of the studies performed diagnosis of multicollinearity for fitting regression models [12]. Similarly, Graham examined 294 ecological papers from 1993 to 1999 and found that only 11% mentioned the possibility of multicollinearity, and about half of them actually tested the presence of multicollinearity [11]. The lack of multicollinearity diagnosis in previous studies

should call our concern, that the health effects of exposures are doubtful if multicollinearity is present in the regression models.

2.4 Conventional Strategies for Solving Multicollinearity

Several strategies have been proposed in the literature to overcome the issue of multicollinearity. Multicollinearity may result from a small sample size or homogenous data that cannot well represent the population. In this case, the problem can be solved by including more data with a larger random sample [12]. However, in practice, additional data is costly and not always available especially for retrospective studies. On the other hand, many explanatory variables are naturally correlated, such as the multiple air pollutants, in which situation the method of increasing sample size is not very effective for solving multicollinearity. Therefore, statistical techniques for resolving the issue of multicollinearity is crucial for achieving a more accurate parameter estimation and interpretability of regression coefficients.

Among the various approaches, the easiest way to tackle multicollinearity is to drop one or some of the correlated explanatory variables, which can be achieved by using model selection methods, such as forward, backward or stepwise selection [11, 21, 22]. Backward model selection is often preferred than the forward model selection, especially when multicollinearity is present. However, the selection of explanatory variables included in the analysis is difficult, as the selection is decided by the importance of each variable, which may be, in turn, influenced by the multicollinearity [21, 22]. The reduction of explanatory variables may also lead to inferential problems. Firstly, the unique contribution of the excluded variables lacks investigation, and the explanatory power may decrease [11]. Secondly, some explanatory variables which are logically related to the outcomes are possibly omitted after the statistical selection. Dropping variables in the regression models thus is not most ideal for solving multicollinearity, although it takes less effort.

PCA is a commonly used alternative approach to solve multicollinearity by substituting the explanatory variables with new variables created based on them [23]. It takes all the correlated factors and the hidden relationships between them into account, by compressing

the high-dimensional data to factors without linear correlation. In PCA, the vectors that can account for the greatest variation in the correlated explanatory variables are identified to create a new set of variables called "principal components" [11]. Since the vectors are orthogonal, multicollinearity is no longer an issue in the modified regression model. However, as principal components are only mathematical values that keep the maximum variance of the original explanatory variables, the interpretation of the coefficients of principal components can be very challenging [13].

Besides the model selection or the PCA method, other methods, such as ridge regression, can also help reduce the impact of multicollinearity on the parameter estimates by decreasing the inflated variance of the parameter estimates. The central idea of the ridge regression is to introduce a small bias in the estimated regression coefficients in exchange for a substantial reduction in the estimated variance. As a result, the estimated coefficients approach to true values much more stably [24]. However, it is always undesirable to have biased parameter estimates, especially when the research interest focuses on studying the covariate effect. Besides, even though ridge regression is adopted, it is still controversial to what degree of bias involved in the regression estimation is acceptable [25].

In environmental health research, a composite air quality index is often developed to evaluate the joint effect of multiple pollutants [15]. The indicator commonly used is called air quality index (AQI), which is created after the air pollutants being rescaled and combined. Then, AQI is included in the health model for evaluating the overall impact of multiple pollutants on the health outcome [15]. As the correlated pollutants are not simultaneously included in the same health model, the problem of multicollinearity no longer exists. AQI focuses on the joint effect of multiple pollutants; however, the interests often lie in studying the marginal effect of each pollutant, which cannot be derived based on the composite score method.

2.4.1 Variable Selection

Variable selection is widely used to deal with multicollinearity due to its simplicity in practice. It was developed in the early 1960s to search for the "best" subset of explanatory variables [26]. In the searching procedure, a sequence of regression models is developed, and depending

on adding or deleting an explanatory variable in each regression model in sequence, the searching procedure can be classified into three general types: forward selection, stepwise regression and backward elimination.

The original form of forward selection decides whether a predictor is included in the final model via a cutoff of p-value, which usually takes 0.05. The procedure begins with a regression model with no predictors. Then the hypothesis on the coefficient $H_0 : \beta_i = 0$ is tested for each i^{th} predictor X_i, and the one with the smallest p-value is added into the model. Following the same manner, the predictors with the smallest p-value in the rest predictors are added in sequence, until there is no longer any p-value below the cutoff value [25].

Stepwise regression provides a modified searching procedure based on forward selection [19]. It also begins with no predictors in the model, but after each following selection step, all the predictors in the current model are re-evaluated. Thus, a new cutoff value is introduced to kick out the predictor of which the p-value is larger than it in the current model. As multicollinearity can render a predictor to be insignificant even it was important in an earlier stage, during the procedure of stepwise regression, a predictor might be deleted when multicollinearity exists after the new explanatory variable is added into the model. Same with forward selection, the searching procedure of stepwise regression terminates when no more predictors have small enough p-values to enter the model.

Backward elimination is an opposite search procedure of the forward selection. It begins with all the possible predictors contained in a regression model, and the predictors with the largest p-values are deleted one by one, until all the predictors left in the model have p-values smaller than the cutoff value.

The original forms using p-values to select an optimal subset of predictors are considered out of date by many researchers nowadays, but the idea of adding or deleting predictors step by step is still favored to deal with multicollinearity [25]. Criteria other than the cutoff of p-values are also used to select predictors into the final regression model. In R, the value of AIC is set as default for selection: In the forward selection and stepwise regression, the predictor which will bring the largest drop in AIC is added into the model firstly; and the predictors which have a negligible impact on AIC are deleted from the model in backward elimination procedure.

Compared to the forward selection, backward elimination is more likely to show the implications of models with many predictors [24]. Predictors that are not important individually but collectively can show a substantial impact on the result could be ignored using forward selection. As a result, many analysts prefer backward elimination to find the final model, in case that forward selection may underfit [25].

2.4.2 Principal Component Analysis

The main idea of PCA is to reduce the dimensionality of variables, meanwhile retaining the variation present in the data set as much as possible. A new set of uncorrelated variables, namely the principal components, are transformed from the data set and ordered according to their contribution to the variation. Therefore, the first few of the ordered principal components can preserve most of the variation to represent the original variables [27].

Let matrix $\boldsymbol{X}$ denote the original data set consisting of K random variables and n observations. Let $\boldsymbol{P}$ stand for the matrix of principal components to re-express the original data set $\boldsymbol{X}$. The strategy to get $\boldsymbol{P}$ is to find n new predictors that are linear combinations of the n original ones.

In order to properly perform PCA, we need to centralize the data first by subtracting the mean of each predictor from them, i.e., the average of each column in $\boldsymbol{X}$ is subtracted from the elements in that column, so the matrix $\boldsymbol{X}$, as well as the new predictors, will have a mean zero after this step. The key idea to find the new predictors is from Equation (2.8) where $\boldsymbol{D}$ is the diagonal matrix. We then have

$$\boldsymbol{D} = \boldsymbol{U}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U} = (\boldsymbol{X}\boldsymbol{U})'\boldsymbol{X}\boldsymbol{U}. \tag{2.8}$$

As the j^{th} column of $\boldsymbol{X}\boldsymbol{U}$ is $\boldsymbol{X}$ times the column j of $\boldsymbol{U}$, indicating a linear combination of the columns of $\boldsymbol{X}$, namely the original predictors, each column in $\boldsymbol{X}\boldsymbol{U}$ thus represents a new variable called the principal component of $\boldsymbol{X}$. From Equation (2.8), $\boldsymbol{D}$ is the covariance matrix of the principal components which is diagonal, that means the principal components are uncorrelated [25]. After the eigenvalues are sorted from largest to smallest, and the corresponding eigenvectors in $\boldsymbol{U}$ are sorted, $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{U}$ is therefore a sorted matrix. Among these sorted new predictors, since the last few have very small variance, which will make them

approximately constant being around zero, they can be ignored due to little contribution in predicting the variations of the variables.

3. Methodology

We propose a two-stage shared component model for simultaneously estimating the effects of highly correlated explanatory variables on the health outcome. The first stage is the pollution model for extracting the common and pollutant-specific residual components. The second stage is the disease model to assess the impact of the extracted common and the pollutant-specific residual components from the first stage on the disease risk.

3.1 Model Specification

3.1.1 Stage 1—Pollution Model

In this first stage, a shared component model is developed to model two or more explanatory variables for extracting the common and pollutant-specific residual variations [28, 29, 30, 31].

Suppose the whole study area is partitioned into n regions. Let $\boldsymbol{X}_k = (X_{1k}, \ldots, X_{nk})$ denote the vector of the observed values of the k^{th} pollutant over n regions. The pollutants $\boldsymbol{X}_k$ $(k = 1, \cdots, K)$ can be re-expressed as a combination of $\boldsymbol{b}$ and $\boldsymbol{h}_k$, where b is a common component shared across the multiple pollutants inducing the correlation among the pollutants. After the variable b is subtracted from each pollutant, the residual component of the pollutant can be derived, which is denoted as h_k. In order to find the values of $\boldsymbol{b}$, we can pick one of the highly correlated risk factors, namely $\boldsymbol{X}_1$, as the reference. The Stage 1 model can be expressed as follows,

$$\begin{cases} \boldsymbol{X}_1 = \alpha_{01}\mathbf{1} + \boldsymbol{b} \\ \boldsymbol{X}_k = \alpha_{0k}\mathbf{1} + g_k\boldsymbol{b} + \boldsymbol{h}_k \quad for\ k = 2, \cdots, K\,, \end{cases} \tag{3.1}$$

where α_{0k} denotes the intercept for the k^{th} pollutant and $\mathbf{1} = (1, \ldots, 1)_{n\times 1}$ stands for the intercept term. $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$ is a vector of the common factor across n regions and

$\boldsymbol{h}_k = (h_{1k}, \ldots, h_{nk})^\top$ is a vector of the residual for the k^{th} pollutant. We assume

$$\begin{cases} \boldsymbol{b} \sim N(0, \sigma_b^2) \\ \boldsymbol{h}_k \sim N(0, \sigma_{h_k^2}) \;\; for\; k = 2, \cdots, K\,. \end{cases} \tag{3.2}$$

where the parameters σ_b^2 and $\sigma_{h_k}^2$ are the variance parameters for the common and residual components, respectively. The parameter g_k is the factor loading parameter of $\boldsymbol{b}$ for $\boldsymbol{X}_k$ and $g_1 = 1$ for $\boldsymbol{X}_1$ to ensure the identifiability of the model [28, 29, 30, 31].

Based on the equation (3.1) $\boldsymbol{b}$ and $\boldsymbol{h}_k$ can be computed as

$$\begin{cases} \boldsymbol{b} = \boldsymbol{X}_1 - \alpha_{01}\mathbf{1} \\ \boldsymbol{h}_k = \boldsymbol{X}_k - \alpha_{0k}\mathbf{1} - g_k\boldsymbol{b} \;\; for\; k = 2, \cdots, K\,. \end{cases} \tag{3.3}$$

This formulation can be extended to include a spatially correlated random effect to jointly modelling the multiple spatially correlated pollutants. Nevertheless, by imposing the spatial structure on the random effect terms may obscure the local features of the spatial distribution of the pollutants. To avoid over-smoothing, we consider $\boldsymbol{b}$ and $\boldsymbol{h}_k$, $k = 2, \cdots, K$ follow independent normal distributions in this investigation [28, 29, 30, 31].

3.1.2 Stage 2—Disease Model

In the second stage of the proposed model, the extracted common and pollutant-specific components from the Stage-1 model are included as the covariates for modelling the disease risk.

Let Y_i and E_i denote the observed and expected number of disease cases in the i^{th} region, respectively. A column of ones for the intercept term and the observations of K covariates are expressed as $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_n^\top)^\top$, where for each region $\boldsymbol{X}_i^\top = (X_{i1}, \cdots, X_{iK})$. Suppose the disease outcome follows a count distribution $g(\cdot)$. The disease model is then given by

$$\begin{aligned} Y_i \mid E_i, \theta_i \sim g(\mu_i),\; \mu_i = E_i\theta_i \;\; for\; i = 1, \cdots, n \\ log(\mu_i) = log(E_i) + \boldsymbol{X}_i^\top\boldsymbol{\beta}, \end{aligned} \tag{3.4}$$

where, the log risk is modelled as

$$log(\theta_i) = \boldsymbol{X}_i^\top\boldsymbol{\beta},$$

where θ_i denotes the relative risk of the disease outcome in the i^{th} region, $i = 1, \cdots, n$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_K)$ represents the vector of regression parameters in the model. The most commonly used count distributions $g(\cdot)$ are Poisson and negative binomial distributions [28, 29, 30, 31].

For comparison, other approaches, including the naive model (including all the highly correlated pollutants as covariates in the disease model), variable selection (including the pollutants after backward model selection as covariates in the disease model), and PCA (including the principal components of the pollutants as covariates in the disease model) , are also applied.

3.2 Model Validation and Evaluation

3.2.1 Root Mean Squared Error

Root mean squared error (RMSE) has been frequently used to measure model performance in many research studies. It is the square root of the quadratic mean of differences between the predicted values and the observed values. RMSE of the regression model (2.1) can be mathematically expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \frac{1}{\hat{y}_i - y_i}}{n}}, \tag{3.5}$$

where $\hat{y}_i$ is the predicted value of the response y_i. A smaller RMSE indicates a better overall model fit. RMSE is believed an appropriate way for model evaluation in research, especially when the model errors are normally distributed. Due to the quadratic term in the calculation of RMSE, the larger absolute values of model errors take more weight in RMSE. Therefore, RMSE is sensitive to the outliers, of which the severe ones may need to be removed when calculating RMSE [32].

3.2.2 Akaike's Information Criterion

Akaike's information criterion (AIC) allows comparisons between multiple competing models to decide which model can best explain the outcome. It is meaningless by itself, but when

the AIC values of multiple models are ranked, the smaller value indicates a better model. When the number of observations is large enough, AIC is calculated as:

$$AIC = -2log(L) + 2(K + 1) \tag{3.6}$$

where L is the maximum likelihood of the $K + 1$ estimated parameters (K explanatory variables plus the intercept). AIC will increase by 2 with K increased by 1. Models with differences of AIC less than 2 ($\Delta AIC < 2$) due to one extra parameter are therefore considered to have the same goodness of fit [33, 34].

3.2.3 Residual Plots

For the diagnosis of the linear regression model, one of the most versatile methods is to plot the residuals. As in the linear regression model (2.1), the error term ϵ is supposed to follow a normal distribution with the mean of zero and the constant variance of σ_ϵ^2. We can assess if these assumptions are met by visualizing the residuals. In the plots of the residuals against the predicted values, the points should scatter randomly around zero, indicating the regression does have a linear relationship with homoscedastic noise. The evaluation of normality for the residuals can be realized by a histogram of the residuals or a quantile-quantile (Q-Q) plot where the values of residuals are sorted and plotted versus the expected values if they are drawn from a normal distribution [35].

4. Application: Glasgow Study

4.1 Motivating Study

This thesis is motivated by an environmental health study by Lee et.al. to examine the impact of air pollutants, such as $PM_{2.5}$, PM_{10} and NO_2 on the risk of developing respiratory diseases in Greater Glasgow, Scotland in the year 2011 [16].

The study population was nearly 1.2 million people who lived in the city of Glasgow and the River Clyde estuary, which comprised 271 administrative units with about 4000 people in each region on average. The response variable was the counts of hospital admissions with diagnosed respiratory diseases (codes J00-J99 and R09.1 of the International Classification of Disease Tenth revision), of which the distribution is shown in Figure 4.1. In order to adjust for the different sizes and demographic structures of populations in each region, the expected counts of hospital admissions were calculated using the external standardization based on age and sex. The standardized incidence ratio, i.e., the ratio of the observed counts to the expected counts of hospital admissions, of each region is displayed in Figure 4.2. The highest risks of respiratory diseases are shown in the middle and eastern study regions. The pollution concentrations were yearly averaged by the authors using dispersion models with the source from the Department for the Environment, Food and Rural Affairs (DEFRA). The median value of the modelled average concentrations at a resolution of 1km grid squares in 2010 in each region was used as the measure of explanatory variables. The pollution concentrations in the year prior to the observed hospital admissions were used to ensure the causal relationship that air pollution exposures should occur before the disease diagnosis [16].

The original aim of the study by Lee et.al. was to propose a localized conditional autoregressive model for modelling the local spatial dependence of the data, and the air pollutants entered the model separately to circumvent the issue of multicollinearity. The study found $PM_{2.5}$, PM_{10} and NO_2 were related to the increased risks of respiratory diseases individually in the studied area [16].

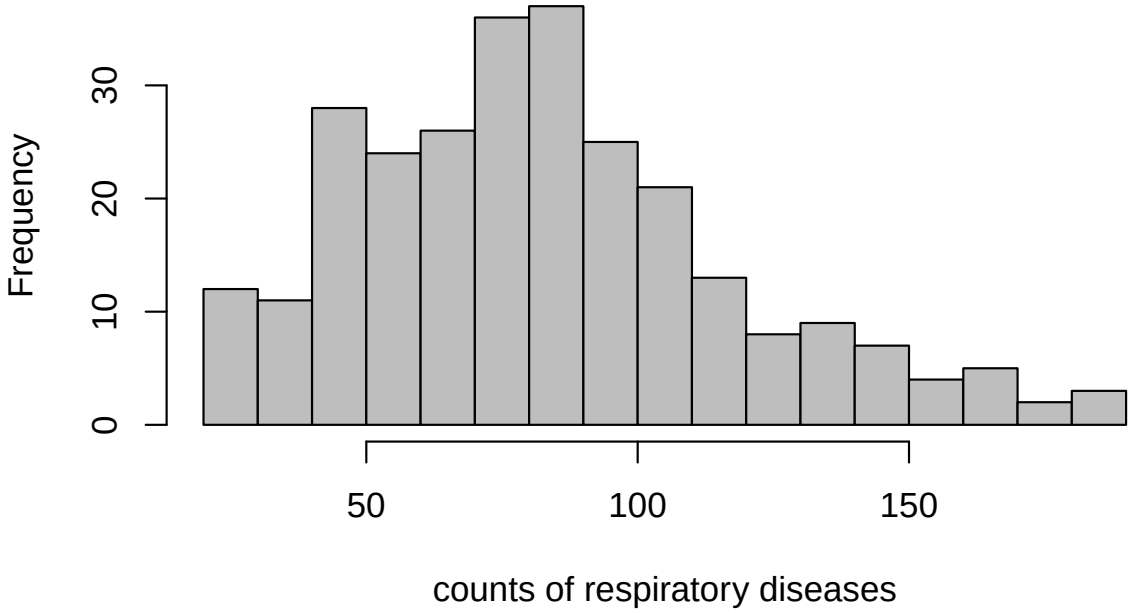


Figure 4.1: The distribution of the observed number of hospital admissions for respiratory diseases over the 271 administrative units in Greater Glasgow, Scotland, in the year 2011.

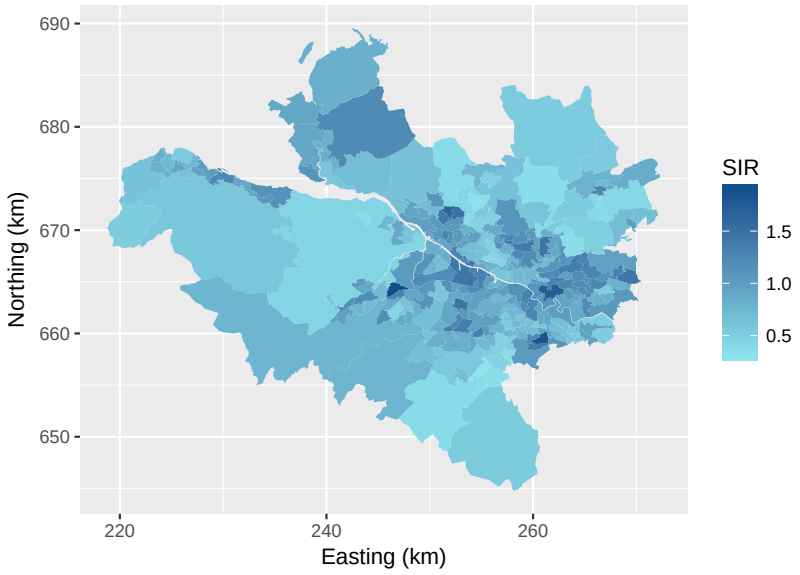


Figure 4.2: Map of the standardized incidence ratio (SIR) of hospital admissions for respiratory diseases over the 271 administrative units in Greater Glasgow, Scotland, in the year 2011.

The goal of this study is not to investigate the methods of modelling spatial correlation of the disease incidence, but rather to develop a modelling method for simultaneously modelling the impact of highly correlated air pollutants on the disease incidence. This data is publicly available from the paper by Lee et.al. [16] at the data repository of the Biometrics website. The ethics exemption from the Biomedical Research Ethics Board at the University of Saskatchewan was attached in Appendix C.

4.2 Exploratory Data Analysis

As shown in Figure 4.3, $PM_{2.5}$, PM_{10} and NO_2 are right-skewed. As a result, we take the square root transformation of the scaled explanatory variables to normalize the distributions. After variable transformations, $PM_{2.5}$, PM_{10} and NO_2 are approximately symmetric (Figure 4.4), with each unit of the transformed values are 1.1, 1.3 and 2.5 $(\mu g m^{-3})^{\frac{1}{2}}$, respectively. All the pollutants are highly and nearly linearly correlated. In particular, $PM_{2.5}$ and PM_{10} are almost perfectly linearly related, with the Pearson correlation coefficient equal to 0.994. NO_2 is also highly related to $PM_{2.5}$ and PM_{10} with the Pearson correlation coefficients equal to 0.983 and 0.974, respectively.

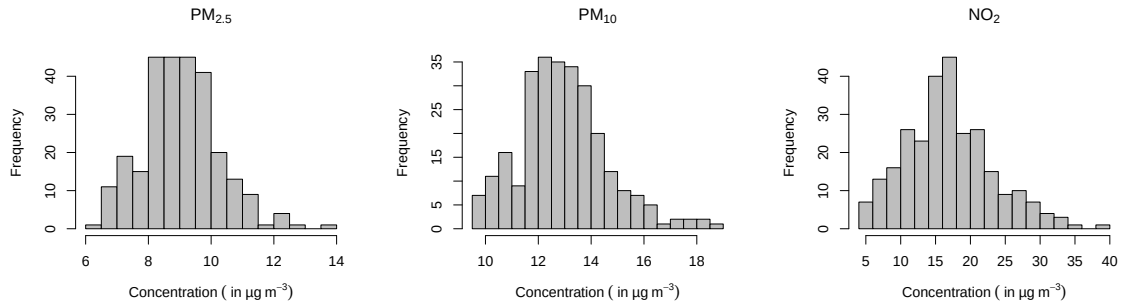


Figure 4.3: The histogram of the yearly concentrations of the three pollutants over the 271 administrative units in Greater Glasgow, Scotland, in the year 2010.

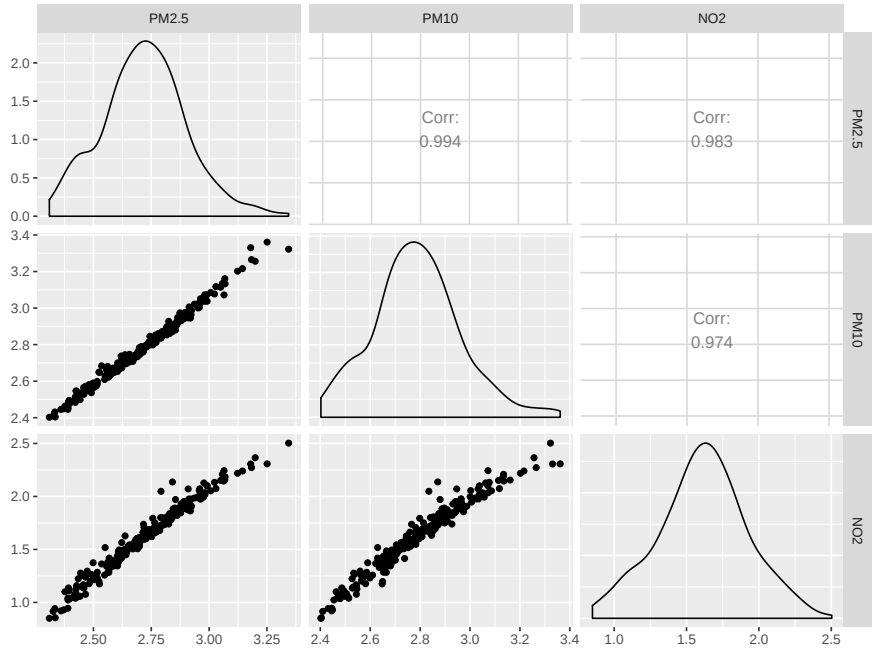


Figure 4.4: The correlations between scaled $PM_{2.5}$, PM_{10} and NO_2 after square root transformation.

4.3 Model Formulation

4.3.1 Stage 1 - Pollution Model

Let X_{i1}, X_{i2}, X_{i3} denote the observed values of $PM_{2.5}$, PM_{10} and NO_2 in the i^{th} region of the study sample. The Stage-1 pollution model can be expressed as:

$$\begin{cases} f(X_{i1}) = \alpha_{01} + b_i \\ f(X_{i2}) = \alpha_{02} + g_2 \cdot b_i + h_{i2} \\ f(X_{i3}) = \alpha_{03} + g_3 \cdot b_i + h_{i3}, \end{cases} \tag{4.1}$$

where $f(\cdot)$ refers to the transformation function of the outcome variables for normalization. In this study, $f(\cdot)$ is the square root transformation. b_i is the shared component of $PM_{2.5}$, PM_{10}, NO_2; h_2 and h_3 represent the residual effects of PM_{10} and NO_2 respectively after accounting for the effect of $PM_{2.5}$. In this study, $PM_{2.5}$ has the strongest linear relationship with the other two pollutants. As a result, the shard component based on $PM_{2.5}$ is expected to best represent the correlation between the three pollutants.

The values of b_i, h_{i2} and h_{i3} in the i^{th} region can be then derived as follows, which will

be included in the Stage-2 disease model:

$$\begin{cases} b_i = f(X_{i1}) - \alpha_{01} \\ h_{i2} = f(X_{i2}) - \alpha_{02} - g_2 \cdot b_i \\ h_{i3} = f(X_{i3}) - \alpha_{03} - g_3 \cdot b_i. \end{cases} \tag{4.2}$$

4.3.2 Stage 2 - Disease Model

Let Y_i and E_i denote the observed and expected numbers of respiratory disease cases in the i^{th} region, respectively.

$$\begin{aligned} Y_i &\sim Poisson(E_i\theta_i), \\ \log(\theta_i) &= \boldsymbol{X}_i^\top \boldsymbol{\beta} + \gamma JSA_i, \end{aligned} \tag{4.3}$$

where θ_i represents the relative risk of the disease in the i^{th} region. The covariates $\boldsymbol{X}_i^\top = (1, X_{i1}, \cdots, X_{iK})$ refer to the scaled pollutants after transformation, or the principal components, or the latent variables based on our proposed method in the Stage-1 pollution model depending on which method is used to fit data. The number of covariates denoted as K may vary depending on which method is used for extracting the information for the correlated pollutants. JSA represents the socio-economic deprivation on health, which was considered as an important confounding variable with its coefficient denoted as γ, to distinguish its effect from the parameters (βs) which we intend to compare.

The Poisson regression model (4.3) assumes equal-dispersion, i.e. $E(Y_i) = Var(Y_i)$. When Y_i is overdispersed ($Var(Y_i) > E(Y_i)$), namely Y_i follows negative binomial distribution, the Poisson regression model (4.3) can be modified as

$$\begin{aligned} Y_i &\sim Poisson(E_i\theta_i), \\ \log(\theta_i) &= \boldsymbol{X}_i^\top \boldsymbol{\beta} + \gamma JSA_i + V_i,\ exp(V_i) \sim Gamma(r, 1/r), \end{aligned} \tag{4.4}$$

where r is referred as the overdispersion parameter, of which the larger value means greater overdispersion [36].

4.4 Results

4.4.1 Stage 1 - Pollution Model

The common and the residual components b, h_2, h_3 for the three pollutants have the mean of zero and the standard deviations 0.182, 0.019, 0.057, respectively. To examine the model fit of the Stage 1 model, residual plots of the models of X_2 and X_3 against the fitted values are produced in Figure 4.5. The points are randomly scattered, so it is reasonable to use the linear models in Stage 1 of our proposed method. The parameters in the models (4.1) have the point estimates as following: $\alpha_{01} = 2.72$, $\alpha_{02} = 2.79$, $\alpha_{03} = 1.61$, $g_2 = 0.97$, $g_3 = 1.66$ (Table 4.1).

The parameters g_2 and g_3 are regression coefficients of b as a covariate for modelling X_2 and X_3. Since the value of b captures the total variability of $PM_{2.5}$, g_2 and g_3 reflect the linear relationships between $PM_{2.5}$ and PM_{10}, $PM_{2.5}$ and NO_2, respectively. $PM_{2.5}$ and PM_{10} both are the measurements of atmospheric particulate matter, so it is not surprising that g_2 is more close to 1 than g_3. The standard deviations of the shared and residual components, b, h_2, h_3 are equal to 0.182, 0.019 and 0.057, respectively, which indicates more variability is captured by the shared component (b_i) compared to the residual components (h_2 and h_3). The stronger correlation between $PM_{2.5}$ and PM_{10} as compared to $PM_{2.5}$ and NO_2 is also reflected in the smaller value of σ_{h_2} than σ_{h_3}, as more residual variability of NO_2 is observed after the common effect is controlled for.

The spatial distributions of $PM_{2.5}$, PM_{10}, and NO_2 are displayed in Figure 4.6. It clearly shows the three pollutants are highly correlated and exhibit very similar geographical pattern, that the highest concentrations tend to be clustered in the middle east part of the study region. As a comparison, Figure 4.7 graphs the spatial distribution of the common and residual components (b and hs) of the explanatory variables across the study region based on the Stage 1-pollution model (shared component model). The pattern of b in Figure 4.7 displays the common effect from the three pollutants, which resembles the spatial patterns of the three pollutants, as shown in Figure 4.6. The spatial map of h_2 does not exhibit any pattern because of little variability in the unique effect of PM_{10} after accounting for the effect

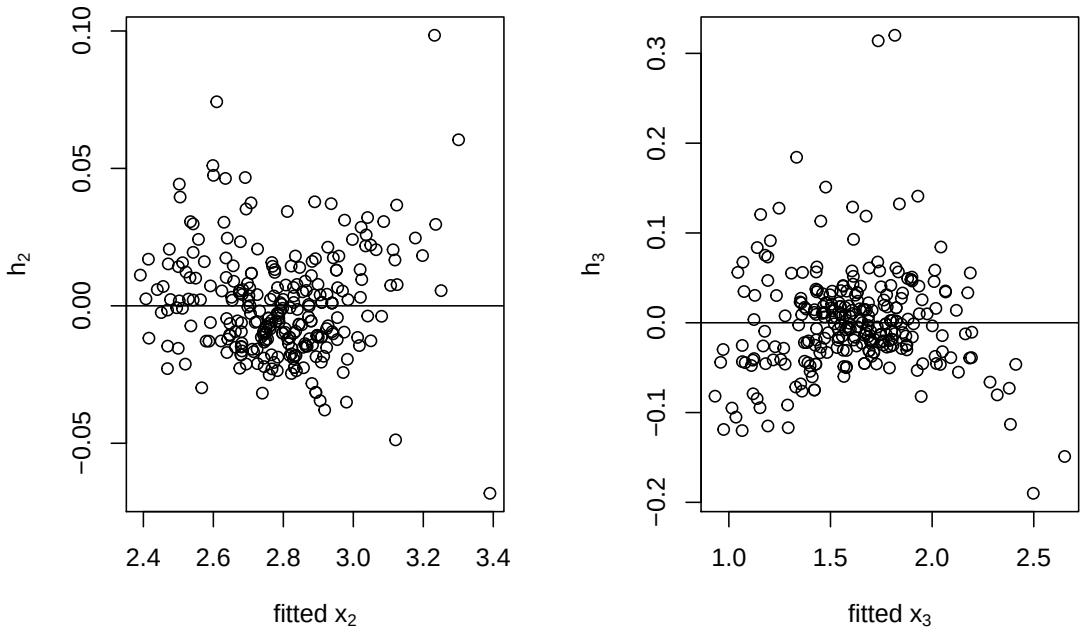


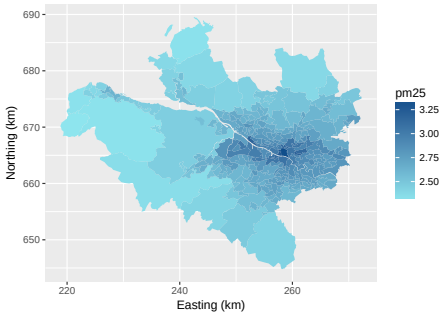
Figure 4.5: The residuals versus fitted values for x_2 (left plot) and x_3 (right plot) based on the State 1-pollution model.

Table 4.1: The estimated regression coefficients and variance components of the shared and residual components of the stage 1 model.

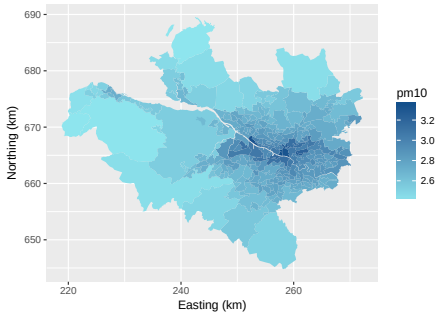
Parameters	Estimate	SE	p-value
α_{01}	2.719	0.0111	< 0.001
α_{02}	2.787	0.00118	< 0.001
α_{03}	1.613	0.00346	< 0.001
g_2	0.966	0.00649	< 0.001
g_3	1.663	0.0190	< 0.001
Parameters	Estimate	DF	
σ_b	0.182	270	
σ_{h_2}	0.019	269	
σ_{h_3}	0.057	269	

SE: standard error
DF: degrees of freedom

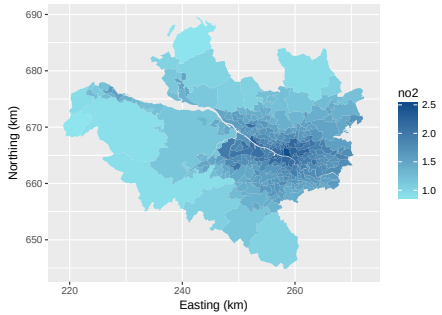
of $PM_{2.5}$. The spatial map of h_3 shows notable residual effects of NO_2 in the middle and the north-western administrative units. Overall, the common effect denoted as b captures the majority of variability from the three pollutants, and little residual effect is attributable to PM_{10} but additional spatial variability is reflected in the residual map of NO_2.



(a) $PM_{2.5}$

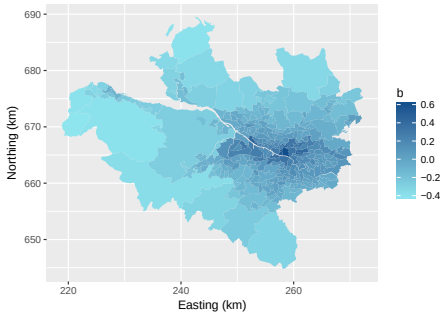


(b) PM_{10}

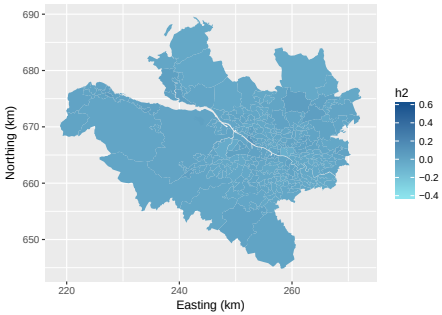


(c) NO_2

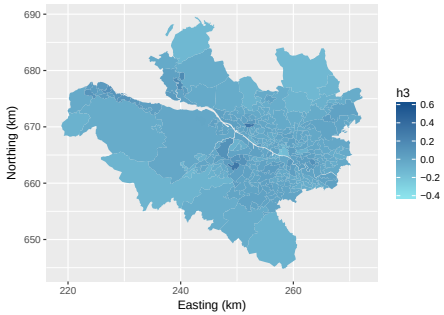
Figure 4.6: Maps of the normalized yearly concentrations of pollutants.



(a) b



(b) h_2



(c) h_3

Figure 4.7: Maps of the shared and residual components of the Stage 1 model.

4.4.2 Stage 2 - Disease Model

Following the first stage of the proposed model where b and hs are obtained, their values are fitted into the disease model to predict the respiratory disease cases. The methods of traditional backward variable selection and PCA are also applied. Below lists all the considered competing methods.

$$\textbf{Model 0:}\quad \log(\theta_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \gamma JSA_i + \upsilon V_i \tag{4.5a}$$

$$\textbf{Model 0-1:}\quad \log(\theta_i) = \beta_0 + \beta_1 X_{i1} + \gamma JSA_i v + \upsilon V_i \tag{4.5b}$$

$$\textbf{Model 0-2:}\quad \log(\theta_i) = \beta_0 + \beta_2 X_{i2} + \gamma JSA_i + \upsilon V_i \tag{4.5c}$$

$$\textbf{Model 0-3:}\quad \log(\theta_i) = \beta_0 + \beta_3 X_{i3} + \gamma JSA_i + \upsilon V_i \tag{4.5d}$$

$$\textbf{Model 1:}\quad \log(\theta_i) = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \gamma JSA_i + \upsilon V_i \tag{4.5e}$$

$$\textbf{Model 2:}\quad \log(\theta_i) = \beta_0 + \beta_1 PC_{i1} + \beta_2 PC_{i2} + \gamma JSA_i + \upsilon V_i \tag{4.5f}$$

$$\textbf{Model 3:}\quad \log(\theta_i) = \beta_0 + \beta_1 b_i + \beta_2 h_{i2} + \beta_3 h_{i3} + \gamma JSA_i + \upsilon V_i \tag{4.5g}$$

$$\textbf{Model 4:}\quad \log(\theta_i) = \beta_0 + \beta_1 b_i + \beta_3 h_{i3} + \gamma JSA_i + \upsilon V_i \tag{4.5h}$$

Model 0 is the naive model, which includes all the three pollutants, following which the Models 0-1, 0-2, 0-3 include only one of the pollutants in each model. Model 1 includes $PM_{2.5}$ and NO_2 after applying the backward variable selection using AIC as the criteria. Model 2 uses PCA method. As shown in Figure 4.8, the first two principal components can explain almost all (99.84%) the variability of the pollution data. As a result, Model 2 includes these two principal components as the explanatory variables in the disease model. Our proposed model (Model 3) includes b and h_2, h_3, as the explanatory variables in the disease model. Model 4 applies a backward selection of b and h_2, h_3, which excludes h_2 from the model.

The listed disease models (4.5) are compared by assuming the count outcome following either the Poisson ($\upsilon = 0$) or the negative binomial model ($\upsilon = 1$) distribution. The estimated parameters for all the fitted models are presented in Table 4.2 (Poisson models) and Table 4.3 (negative binomial models). In both tables, each column presents the relative risk with

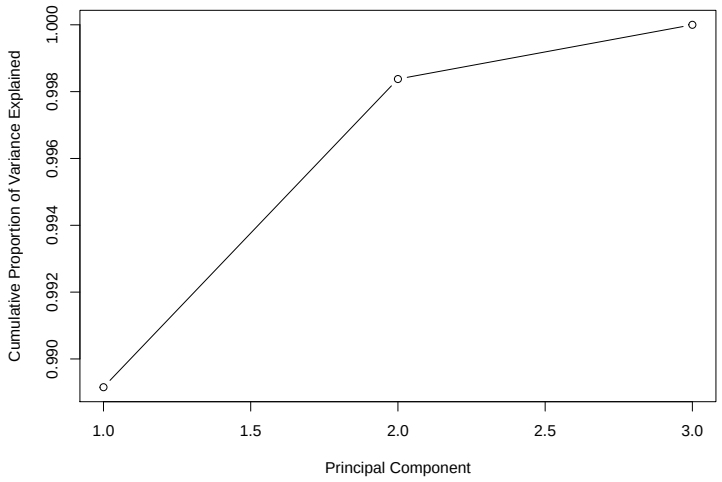


Figure 4.8: The scree plot from PCA.

the 95% confidence intervals (CI) for a one unit increase in the covariate value in that fitted model. The first three rows present the estimated regression coefficients for the air pollutants, the principal components or the shared and residual components, and the row labelled as γ shows the effect of JSA as a confounder. The remaining rows present the values of AIC and -2 log-likelihood, which are used for comparing the overall fits of the models.

As shown in Tables 4.2 or 4.3, the values of the goodness-of-fit tests, i.e., AIC and log-likelihood do not vary much between the models, which is in line with the theory that multicollinearity does not affect the overall model fit. In Figures 4.9 and 4.10, the observed response values are plotted against the fitted response values based on the Poisson and negative binomial models, respectively. As shown in all the plots in both Figures, the points fall along the diagonal line closely, so both the Poisson and the negative binomial models fit the data reasonably well. This confirms that multicollinearity does not affect the overall model fit.

The model fit was further examined by plotting the randomized quantile residuals (RQRs), because our response variable is discrete that the Pearson and deviance residuals will shape some nearly parallel curves, which can distort the messages from the residual plots. Randomization is therefore used to generate RQRs, which are continuously distributed when responses are discrete and consequently, the residual plots can offer meaningful information

on model check [37, 38]. The plots of RQRs against the fitted values of the response variable based on the Poisson and negative binomial models are shown in Figures 4.11 and 4.12, respectively. The plots indicated that RQRs are randomly scattered between -3 and 3 without any discernible pattern. The normality of RQRs is examined by the Q-Q plots. As shown in Figures 4.13 and 4.14, RQRs approximately follow a normal distribution. Hence, the assumptions of the disease models appear to be reasonable. Comparing between the Poisson and the counterpart negative binomial models, the negative binomial models yield smaller AIC and -2 Log-likelihood, which indicates negative binomial models perform better than the Poisson models. Therefore, we will focus on interpreting the parameter estimates based on the negative binomial models.

The results based on the naive model (Model 0) suggest only NO_2 has a statistically significant effect on the respiratory diseases, with 55% higher relative risk relating to its one unit increase when controlling for the other two pollutants. Although the three single-pollutant models all show a strong association between the pollutant and the respiratory disease risk, their overall model fits are slightly worse than the naive model, in particular, when only $PM_{2.5}$ or PM_{10} is included in the model. Model 1 uses the backward model selection to mitigate the problem of multicollinearity. After backward model selection, PM_{10} is eliminated, $PM_{2.5}$ and NO_2 are retained in the model; however, $PM_{2.5}$ has a protective effect on respiratory diseases, which conflicts with the previous findings [1, 2]. The opposite signs of the effects of $PM_{2.5}$ in Model 0 and Model 1 could be attributable to multicollinearity between the explanatory variables. Therefore, dropping PM_{10} after variable selection cannot fully resolve the problem caused by multicollinearity.

PCA is applied in Model 2 that the first two principal components are included as the covariates in the disease model. The results of the disease model indicate that the two principal components are all significantly associated with the disease outcome. However, it is very challenging to interpret how the three pollutants affect the yearly respiratory disease cases based on the protective effects of the principal components.

Model 3 includes the common and residual effects of the pollutants (b, h_2, and h_3) as covariates in the disease model for modelling the respiratory disease risk. The results indicate that for every one unit increase in b, the relative risk of respiratory diseases significantly

increases by 17.4%. PM_{10} does not have a statistically significant residual effect on the disease risk after accounting for the effect of $PM_{2.5}$. By contrast, NO_2 substantially increases the relative risk of respiratory diseases even after accounting for the effect of $PM_{2.5}$, that every one unit increase in the residual component from NO_2 is associated with the relative risk of respiratory diseases increased by 55%. In other words, the impact of NO_2 on respiratory health is greater than the impact of $PM_{2.5}$ in this disease model. After the residual effect of PM_{10} (h_2) is dropped via backward selection, the results from Model 4 show that the shared component b and residual component h_3 exhibit a similar effect on respiratory health to the findings from Model 3, that they are associated with an increased relative risk of 17.3% and 53.1%, respectively. In conclusion, the results from our proposed two-stage shared component model indicate that the shared component of the three pollutants exhibits a significant effect, and NO_2 has an additional effect on respiratory disease risk.

The shared component b is obtained based on $PM_{2.5}$; however, the results may depend on which pollutant is used for extracting the shared component b. In order to investigate if the results of the proposed model are consistent no matter which pollutant enters the proposed model first to decide the value of b, we assign X_1 as PM_{10} or NO_2 to re-run the proposed two-stage shared component model. As shown in Appendix A, the shared component b still significantly increases the respiratory disease risk, regardless of the choice of the pollutant for X_1. The relative risk is equal to 1.227 (95% CI, 1.134 to 1.327) from the Poisson model or 1.180 (95% CI, 1.028 to 1.354) from the negative binomial model when X_1 is PM_{10}, and equal to 1.136 (95%CI, 1.085 to 1.188) from the Poisson model or 1.114 (95% CI, 1.029 to 1.206) from the negative binomial model when X_1 is NO_2 for every unit increase in b. Although the residual components representing the unique effects from the other two pollutants may appear insignificant, even variable selection is applied in the negative binomial models; their trends coincide with the findings above when X_1 is $PM_{2.5}$. Specifically, NO_2 shows stronger effect on the respiratory disease risk compared to the atmospheric particulate pollutants. When PM_{10} enters the model first, every unit of the residual component from NO_2 significantly increase the relative risk by 44.6% in the Poisson model, and when NO_2 enters the model first, one unit of the residual component from $PM_{2.5}$ is associated with the relative risk equal to 0.671 (95% CI, 0.454 to 0.994) in the Poisson model after variable selection.

Table 4.2: A summary of the estimated parameters and the overall fit of the Poisson models.

	Model 0	Model 0-1	Model 0-2	Model 0-3	Model 1	Model 2	Model 3	Model 4
e^{β_1}	0.498 (0.226, 1.098)	1.208* (1.120, 1.302)	-	-	0.671* (0.454, 0.994)	0.979* (0.971, 0.987)	1.219* (1.130, 1.315)	1.216* (1.128, 1.312)
e^{β_2}	1.339 (0.683, 2.620)	-	1.217* (1.126, 1.315)	-	-	0.912* (0.841, 0.988)	1.339 (0.683, 2.620)	-
e^{β_3}	1.446* (1.141, 1.829)	-	-	1.132* (1.082, 1.184)	1.430* (1.131, 1.807)	-	1.446* (1.141, 1.829)	1.430* (1.131, 1.807)
e^{δ}	2.110* (2.033, 2.191)	2.135* (2.059, 2.214)	2.132* (2.055, 2.211)	2.125* (2.049, 2.204)	2.116* (2.040, 2.196)	2.123* (2.047, 2.203)	2.110* (2.033, 2.191)	2.116* (2.040, 2.196)
AIC	2505.8	2511.5	2511.2	2506.5	2504.6	2506.4	2505.8	2504.6
-2LL	2495.9	2505.5	2505.2	2500.5	2496.6	2498.4	2495.9	2496.6

Notes: Model 0: Naive model; Model 0-1, 0-2, 0-3: Single pollutant model; Model 1: Backward selection model; Model 2: PCA model; Model 3: Two-stage shared component model; Model 4: Backward selection of the two-stage shared component model.
The estimated covariate effects are presented as relative risks for one unit increase in each covariates value.

Table 4.3: A summary of the estimated parameters and the overall fit of the negative binomial models.

	Model 0	Model 0-1	Model 0-2	Model 0-3	Model 1	Model 2	Model 3	Model 4
e^{β_1}	0.448 (0.102, 1.957)	1.167* (1.021, 1.334)	-	-	0.578 (0.287, 1.160)	0.983* (0.969, 0.997)	1.174* (1.028, 1.342)	1.173* (1.027, 1.340)
e^{β_2}	1.275 (0.369, 4.417)	-	1.172* (1.021, 1.347)	-	-	0.888* (0.771, 1.022)	1.275 (0.369, 4.417)	-
e^{β_3}	1.550* (1.019, 2.361)	-	-	1.111* (1.027, 1.203)	1.531* (1.012, 2.320)	-	1.550* (1.019, 2.361)	1.531* (1.012, 2.320)
e^{δ}	2.160* (2.020, 2.309)	2.186* (2.048, 2.334)	2.183* (2.045, 2.332)	2.176* (2.038, 2.323)	2.165* (2.028, 2.312)	2.175* (2.037, 2.322)	2.160* (2.020, 2.309)	2.165* (2.028, 2.312)
AIC	2258.2	2258.3	2258.4	2256.7	2256.3	2257.0	2258.2	2256.3
-2LL	2246.2	2250.3	2250.4	2248.7	2246.3	2247.1	2246.2	2246.3

Notes: Model 0: Naive model; Model 0-1, 0-2, 0-3: Single pollutant model; Model 1: Backward selection model; Model 2: PCA model; Model 3: Two-stage shared component model; Model 4: Backward selection of the two-stage shared component model.
The estimated covariate effects are presented as relative risks for one unit increase in each covariates value.

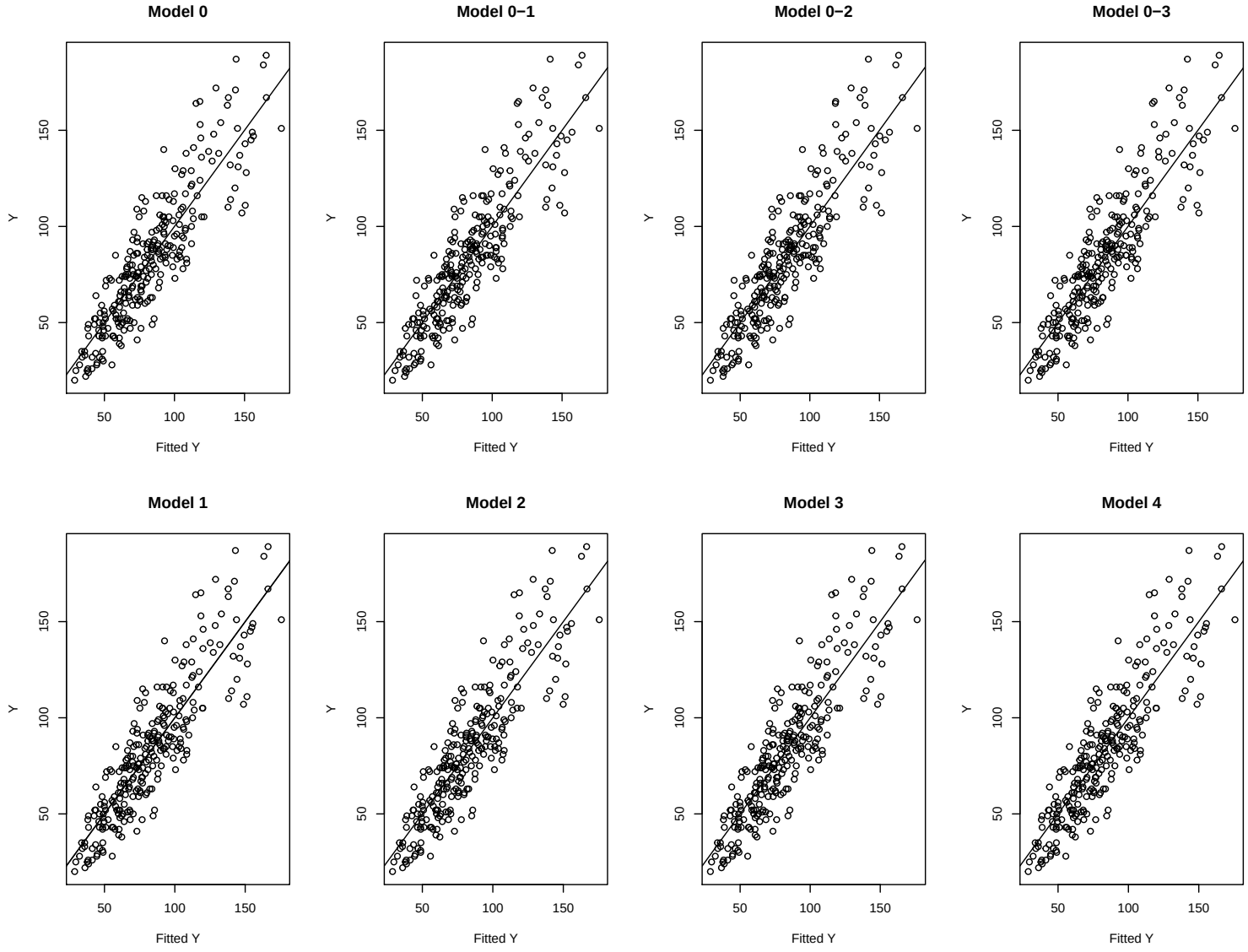


Figure 4.9: The observed disease cases vs. fitted disease cases from the Poisson models.

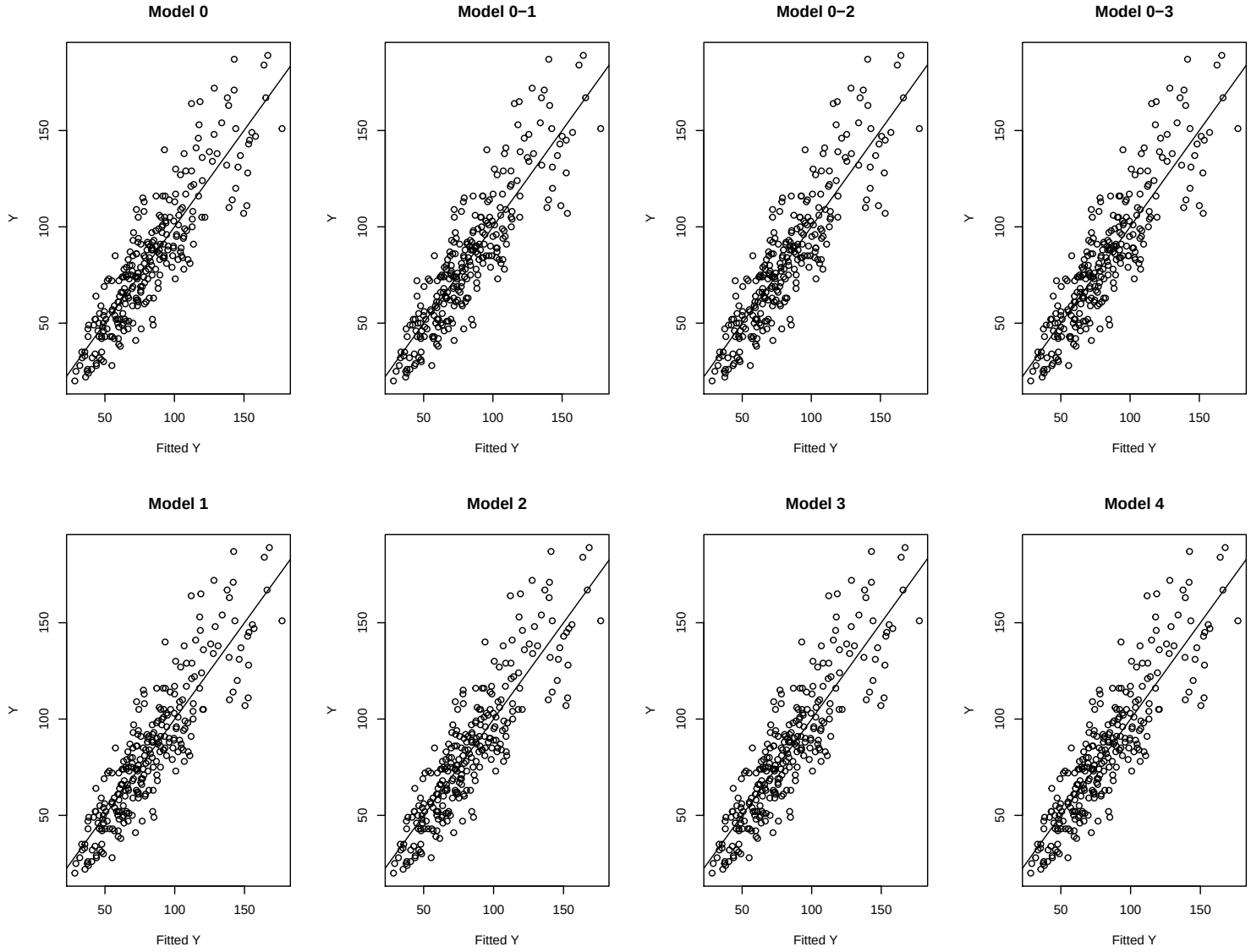


Figure 4.10: The observed disease cases vs. fitted disease cases from the negative binomial models.

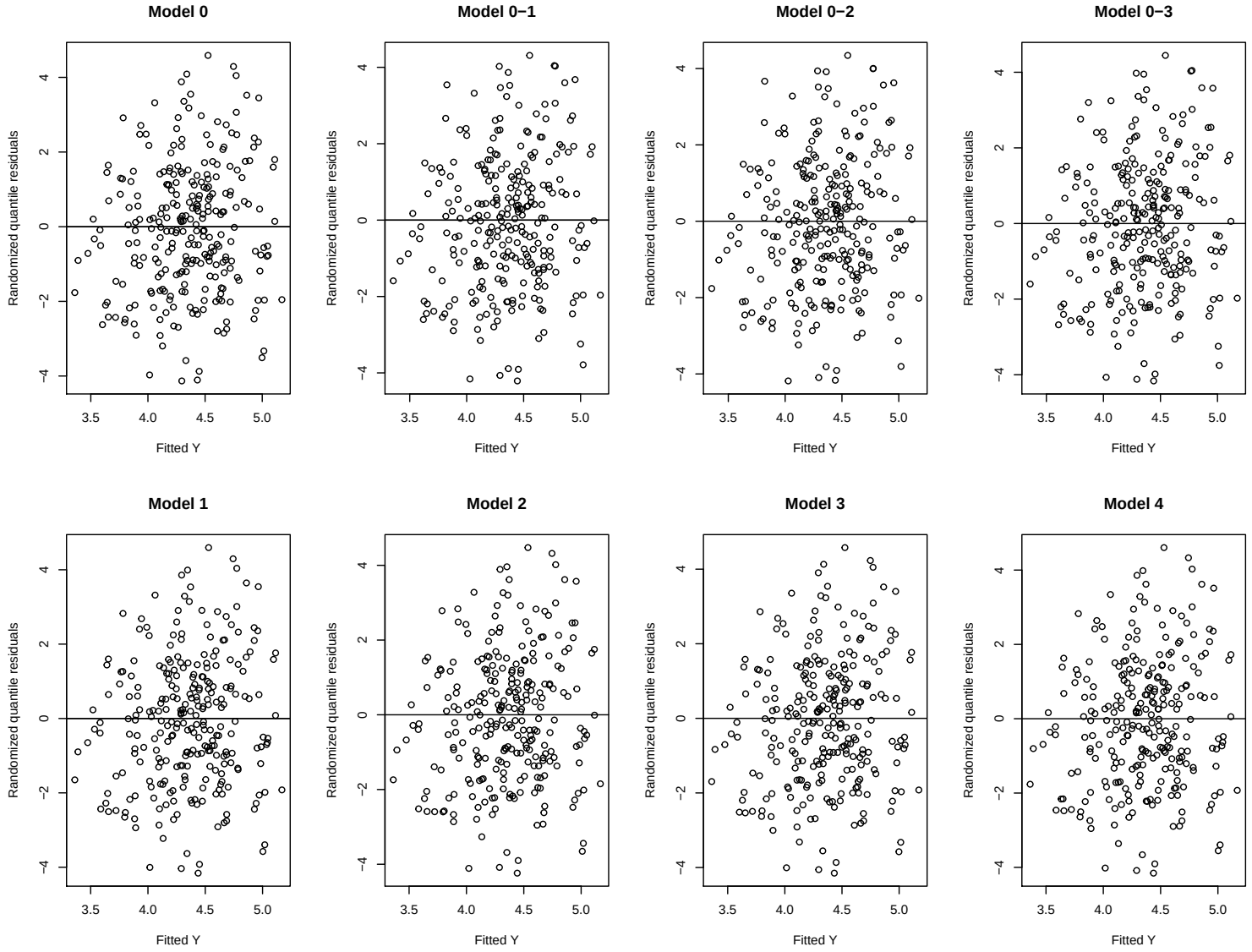


Figure 4.11: Plots of RQR vs. fitted number of disease count from the Poisson models.

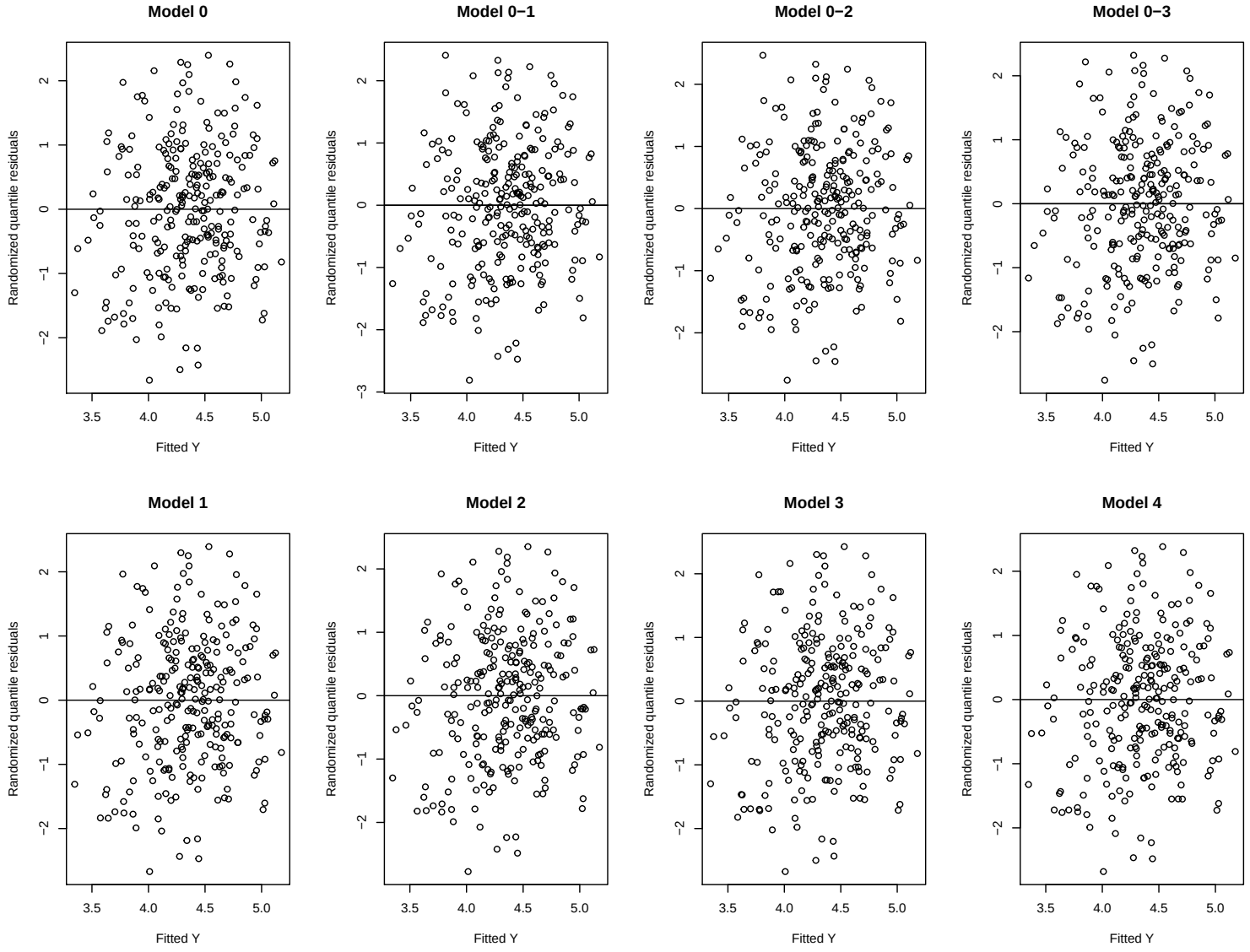


Figure 4.12: Plots of RQR vs. fitted number of disease count from the negative binomial models.

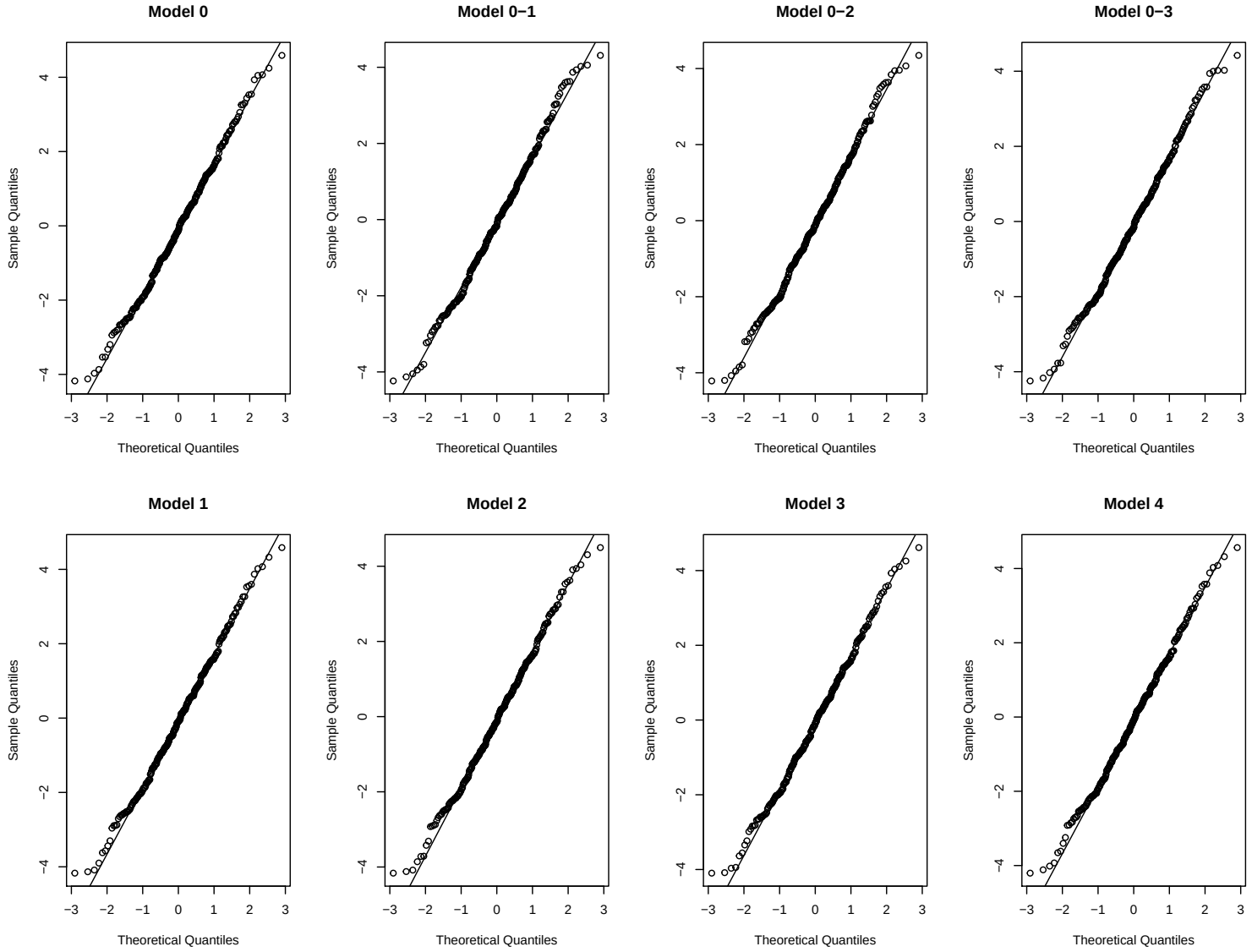


Figure 4.13: The Q-Q plots of RQR for the Poisson models

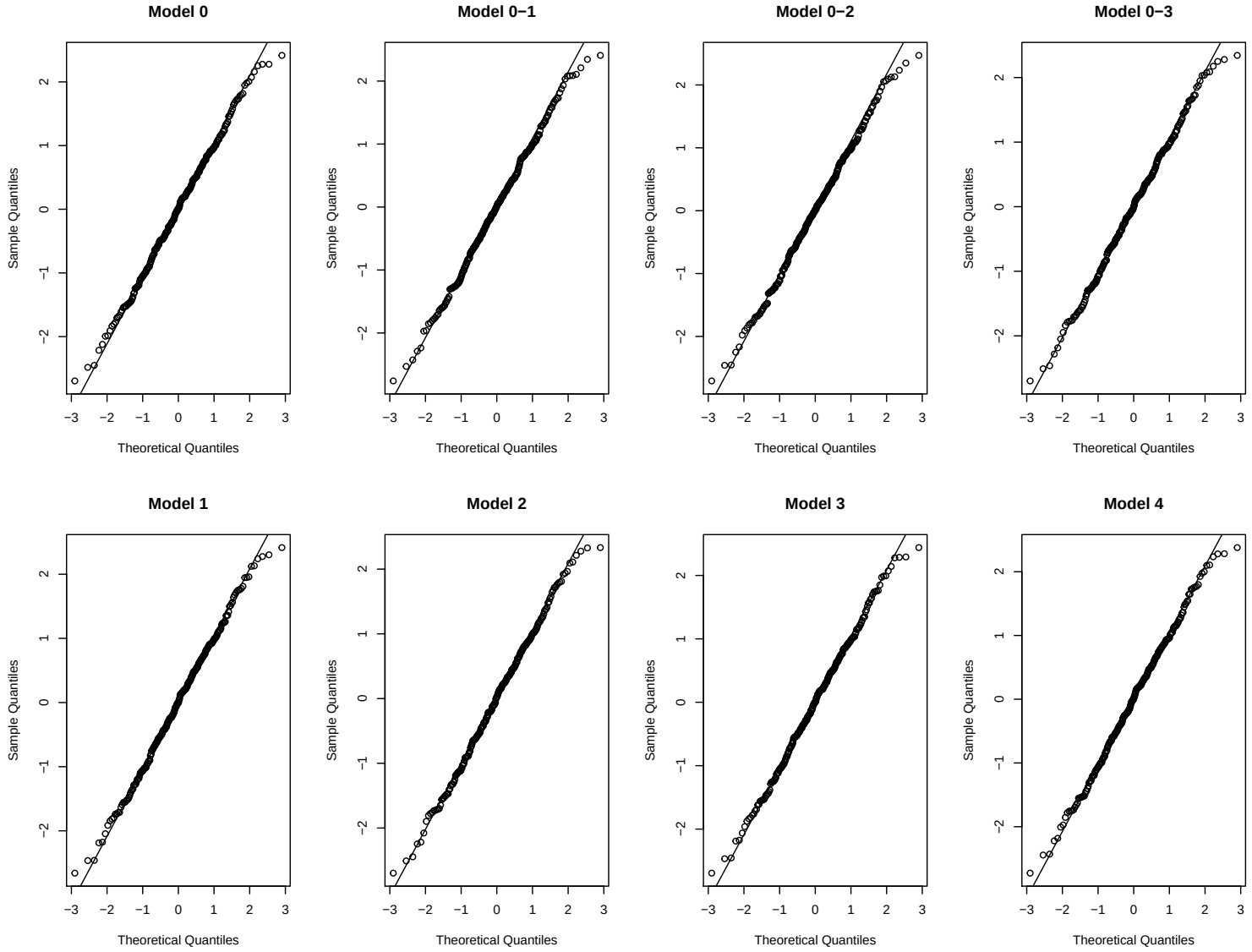


Figure 4.14: The Q-Q plots of RQR for the negative binomial models.

5. Simulation Study

In this chapter, we investigate the properties and performance of the proposed two-stage shared component model against the naive method, including all the correlated explanatory variables in the model, as well as a two-stage model using PCA to extract the latent components.

5.1 Data Generation

Data are repeatedly generated based on a two-stage shared component model of sample size n = 300 for 1000 times. For each simulated dataset, three highly correlated explanatory variables x_1, x_2, x_3 were generated based on a shared component model. The response variable is simulated from a Poisson regression with the expected mean modeled based on the common and residual components of the explanatory variables.

Stage 1 Model (Simulating Explanatory Variables): For each simulated dataset, the common effect b and the residual effects h_2, h_3 are generated from the normal distribution as below:

$$\begin{aligned} b_i &\sim N(0, \sigma_b^2) \\ h_{i2} &\sim N(0, \sigma_{h_2}^2) \\ h_{i3} &\sim N(0, \sigma_{h_3}^2) \\ &for\ \ i = 1, \ldots, n. \end{aligned} \tag{5.1}$$

The three explanatory variables denoted as x_1, x_2, x_3 are then generated as:

$$\begin{cases} x_{i1} = \alpha_{01} + b_i \\ x_{i2} = \alpha_{02} + g_2 \cdot b_i + h_{i2} \\ x_{i3} = \alpha_{03} + g_3 \cdot b_i + h_{i3}, \end{cases} \tag{5.2}$$

where the values of the parameters α_{01}, α_{02} and α_{03}, g_2, g_3 are specified close to the values of the corresponding parameters in the real application. Specifically, we refer to the findings

(Tables 4.1 and 4.2) in the real data analysis and set $\alpha_{01} = 3$, $\alpha_{02} = 3$, $\alpha_{03} = 2$, $g_2 = 1$, $g_3 = 1$.

Based on the equations (5.1) and (5.2), the strength of the correlation among the explanatory variables can be captured by the variability of b in comparison to the variance of the residual components. More specifically, the proportion of variability explained by b, denoted as ψ can be expressed as,

$$\begin{aligned} \psi &= \frac{g^2\sigma_b^2}{g^2\sigma_b^2 + \sigma_h^2} \\ &= \frac{g^2\frac{\sigma_b^2}{\sigma_h^2}}{g^2\frac{\sigma_b^2}{\sigma_h^2} + 1}, \end{aligned} \tag{5.3}$$

which indicates that the magnitude of the correlation between the explanatory variables is determined by the ratio of σ_b to σ_h and the value of g. The values of $\sigma_b/\sigma_{h_3} = 5$ and $\sigma_b = 0.2$, g=1 are close to the corresponding parameter estimates in the real data analysis. To investigate the performance of the proposed method at varying degree of multicollinearity among the covariates, we increase the ratio $\sigma_b/\sigma_{h_3} = 2$, 5, and 10 and $\sigma_b = 0.2$, 0.4, 0.6, 0.8, and 1.0. We hold g=1, since g and σ_b/σ_{h_3} have a similar impact on the correlation among the covariates.

The degree of multicollinearity among the simulated explanatory variables can be measured by the correlation coefficient (r) and VIF. Table 5.1 presents the mean values of r between three simulated explanatory variables and VIF of them over the 1000 simulated samples by setting the ratio of σ_b to σ_{h_3} at 2, 5, and 10, respectively. The results confirm the simulated explanatory variables are highly correlated with r ranging from 0.89 to almost 1 at an increased value of σ_b/σ_{h_3}. By Equation (5.3), a larger σ_b/σ_{h_3} leads to a higher correlation among the covariates, as more variability in the explanatory variables is explained by the common component, b.

Stage 2 Model (Simulating Response Variable): The response variable is generated from a Poisson distribution with the logarithm of its mean as a linear combination of latent variables b, h_2 and h_3 generated based on the Stage 1 model:

$$\log(\mu_i) = \beta_0 + \beta_1 b_i + \beta_2 h_{i2} + \beta_3 h_{i3} \tag{5.4}$$

Table 5.1: The average of the estimated correlation coefficient (r) and variance inflation factor (VIF) for the explanatory variables from the 1000 simulated samples.

σ_b/σ_{h3}	r_{12}	r_{13}	r_{23}	VIF_1	VIF_2	VIF_3
2	0.995	0.894	0.890	107	102	5
5	0.995	0.981	0.976	128	102	26
10	0.995	0.995	0.990	204	102	102

where the regression coefficients are set as β_0=0, β_1=0.5, β_2=0.2, β_3=0.8 to mimic the results based on the real data application. Then, three models, including the naive model, two-stage PCA, and our proposed two-stage shared component model, were fitted to the simulated datasets.

5.2 Results

The performances of the proposed two-stage shared component model in comparison to other competing methods are evaluated by comparing the parameter estimates, their standard errors, and the power of detecting the covariate effect. The goodness-of-fit for all the methods is evaluated according to AIC and RMSE.

The mean values of the estimated regression coefficients over the 1000 simulated datasets are shown in Figure 5.1. The plots in the top panel of Figure 5.1 present the mean values of the estimated regression coefficients over the 1000 simulated datasets in the stage 2 model of the two-stage shared component model. The results indicate that the estimation of the effect of b is very close to its true value. In contrast, as shown in the first plot in the second row of Figure 5.1, the estimated regression coefficient for x_1 based on the naive model tends to be negative. The coefficient estimates of x_2 and x_3 from the naive model are nearly the same as the estimated regression coefficients for h_2 and h_3 from the proposed two-stage shared component method. The results indicate that the inclusion of multiple highly correlated explanatory variables in the regression model can lead to a misleading interpretation of the covariate effects, especially for the covariate x_1. In contrast, the residual effects of x_2 and x_3 are not very much affected by multicollinearity. The bottom panel of Figure 5.1 suggests

that regression coefficients for the extracted PC_1 and PC_2 are around 0. The results show that although the PCA method can reduce the dimensionality of the explanatory variable, it is challenging to find a reasonable interpretation of the effects of the extracted principal components on the disease risk.

Figure 5.2 presents the average values of the standard errors of the estimated regression coefficients over the 1000 simulated samples based on the three models. Due to multicollinearity, the standard error of the estimated regression coefficient for x_1 is inflated. Despite its decreasing trend as σ_b increases, the standard errors of the estimated regression coefficient for x_1 are consistently higher than the values for the b based on the proposed two-stage shared component model. The estimated standard errors for h_2 and h_3 based on the two-stage shared component model are very similar to the estimated standard errors for x_2 and x_3 based on the naive method. The PCA method yield very low standard errors of the estimated effect of PC_1 and PC_2.

The increased standard errors for the correlated covariates also result in decreased statistical power to detect significant explanatory variables. This is reflected in Figure 5.3, which displays the probability of identifying significant regression coefficients at the 5% significance level among the 1000 samples. For the naive method, the statistical power of detecting the significant effect of x_1 is very low under various scenarios. In contrast, the power of detecting the significant effect of b increases as the σ_b increases. More specifically, when σ_b is set to a smaller value, such as 0.2, the estimated covariate effect of b appears to be insignificant, since the variability of common effect (σ_b) built based on x_1 is very low. However, as σ_b increases to 0.4, the power of detecting the significant effect of the common component increases drastically. The statistical power of detecting the effect of b also does not depend on the ratio of σ_b/σ_{h_3}. The statistical power of detecting the effect of h_2 is consistently very low, i.e., around zero, which is not surprising, since the true value of the effect of h_2 was set at a very small value to mimic the real data application. The power of detecting the effect of h_3 increases when σ_b increases, in particular, when σ_b/σ_{h3} is at a smaller value. The high statistical power of estimating the effect of principal components based on the two-stage PCA method indicates that the PCA method is a helpful strategy for tackling the issue of multicollinearity among covariates. Nevertheless, the interpretation of the effects of principal

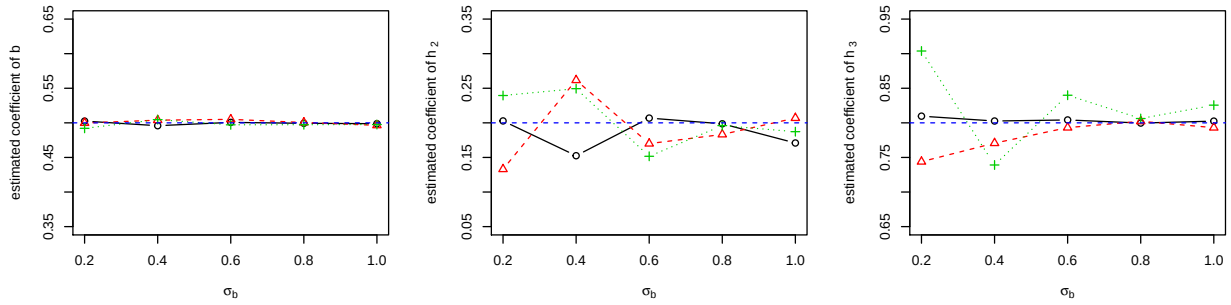
components is challenging, despite their significant effects.

To examine the overall model fit of the three modelling methods, Figures 5.4 and 5.5 present the average RMSE of the estimated response variable and the average AIC of the models, which demonstrate that the overall fits of the models are not affected by multicollinearity.

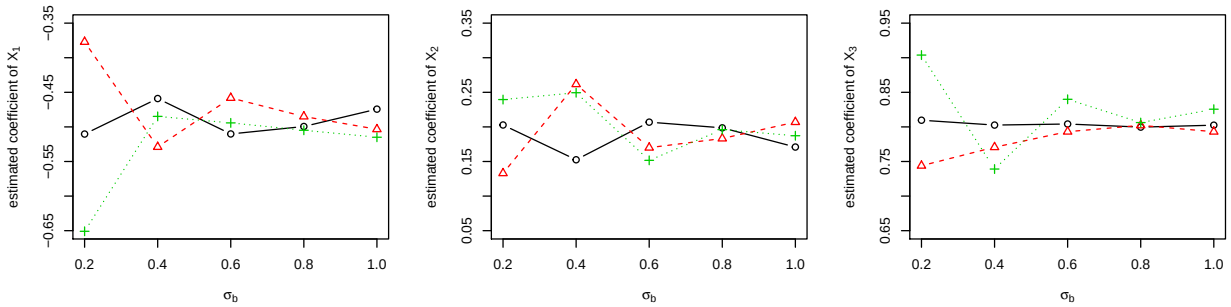
In summary, the proposed two-stage shared component model yield almost unbiased parameter estimates. Under the naive method, multicollinearity among the covariates has a severe impact on the estimated regression coefficient for x_1. Although the PCA method can mitigate the impact of multicollinearity on estimating the covariate effects, the interpretation of the coefficients is very challenging.

In this motivating example, the three pollutants are all positively correlated; however, in some cases, the explanatory variable might be highly and negatively correlated. To investigate the robustness of the proposed two-stage shared component model, an additional set of simulated study was carried out, where the explanatory variables are negatively correlated, as demonstrated by the negative correlation coefficients r_{13} and r_{23} (Appendix B.1). The setting for this additional simulation study is very similar to the previous setting for positively correlated exposures. As shown in the figures B.1, B.2, B.3, B.4, B.5 in Appendix B, the results give consistent conclusions as the ones when the explanatory variables are positively correlated.

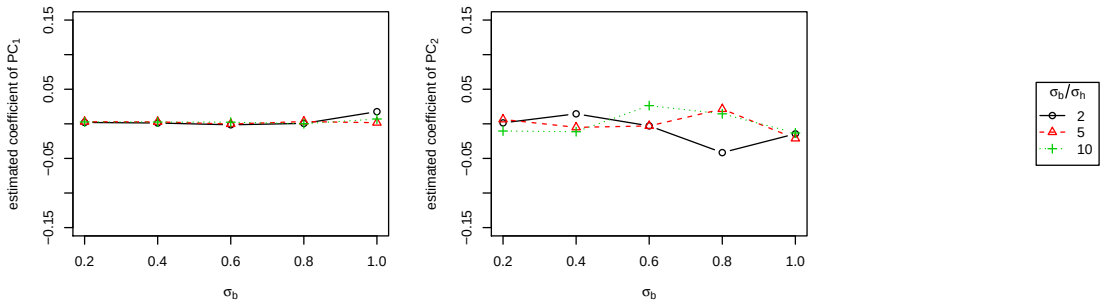
Moreover, in reality, the explanatory variables may have a negative effect on the outcome, so we conducted additional simulation study by setting the regression coefficient, i.e., β_1, as a negative value for a further investigation on the model performance. Consistent results are obtained as well. In conclusion, the direction of the relationship between explanatory variables or the direction of the relationship between explanatory variable and outcome has little impact on model performance.



(a) The average of the estimated regression coefficients of the proposed two-stage shared component models, i.e., the Poisson regression including the common and residual effects of the explanatory variables.

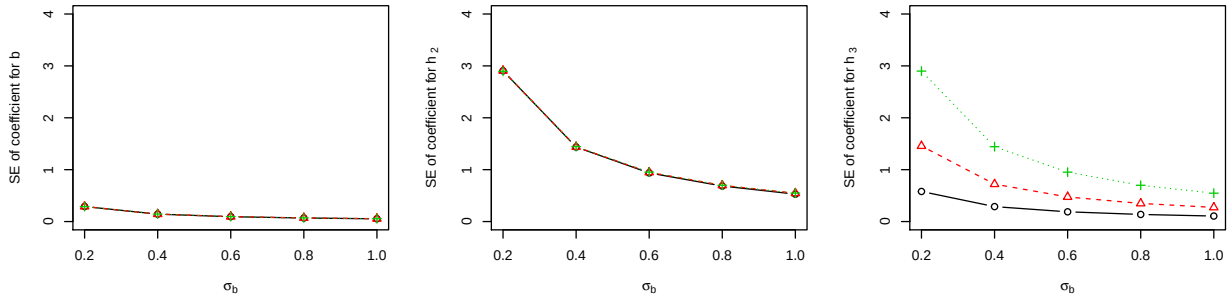


(b) The average of the estimated regression coefficients of the naive model, i.e., the Poisson regression including all the explanatory variables in the model.

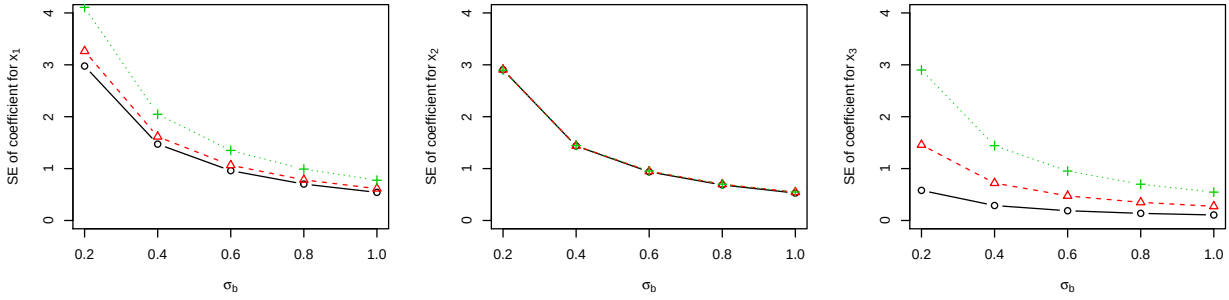


(c) The average of the estimated regression coefficients of the two-stage PCA model, i.e., the Poisson regression including the principal components of the explanatory variables.

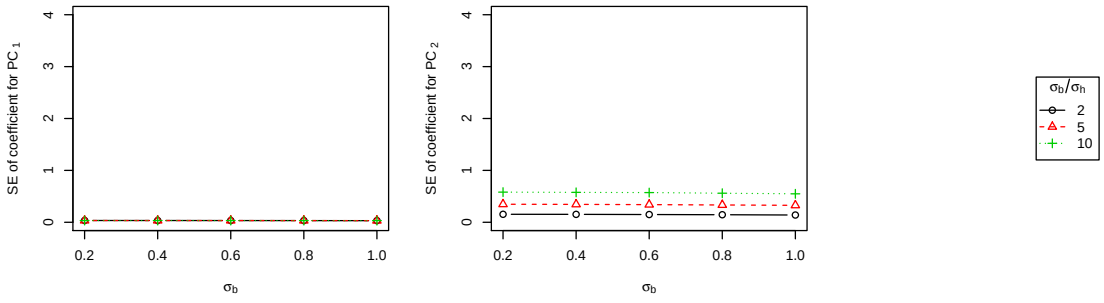
Figure 5.1: The average of the estimated regression coefficients of the proposed two-stage shared component model (top panel), the naive model (middle panel) and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.



(a) The average of the standard errors of the estimated regression coefficients for the two-stage shared component model.

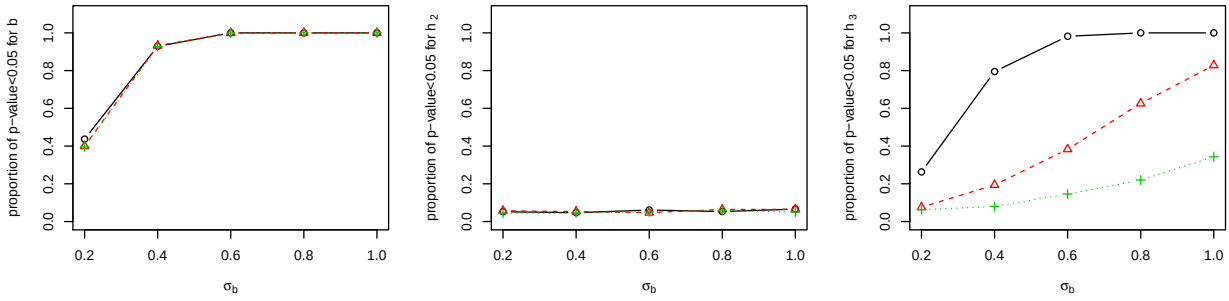


(b) The average of the standard errors of the estimated regression coefficients for the naive model.

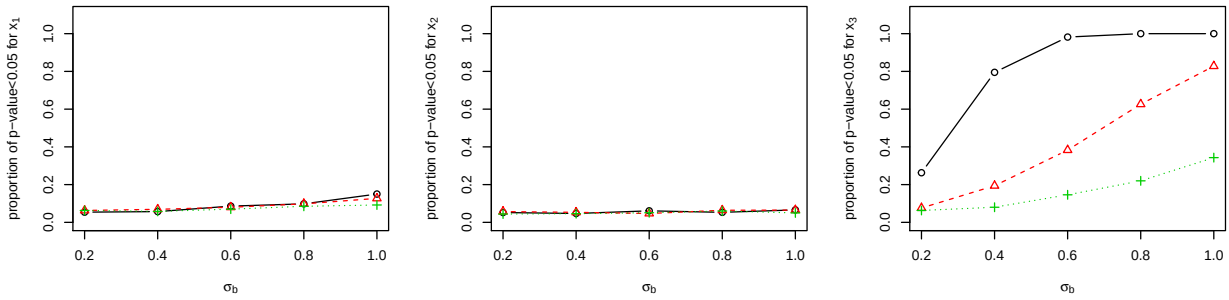


(c) The average of the standard errors of the estimated regression coefficients for the two-stage PCA model.

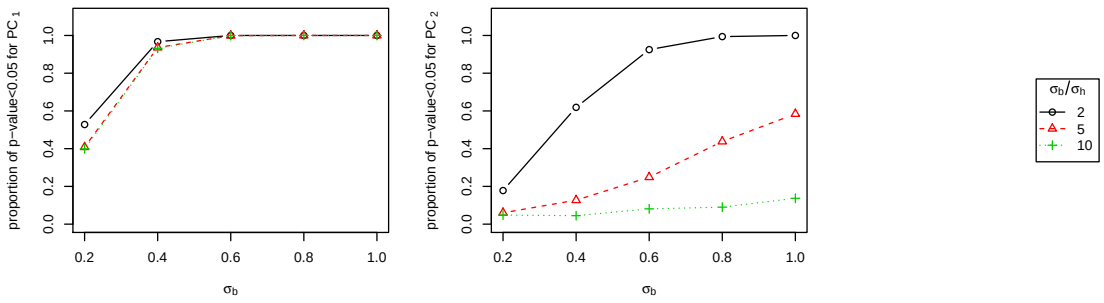
Figure 5.2: The average of the standard errors of the regression coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.



(a) The probability of statistically significant coefficients for the two-stage shared component model.



(b) The probability of statistically significant coefficients for the naive model.



(c) The probability of statistically significant coefficients for the two-stage PCA model

Figure 5.3: The probability of statistically significant coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.

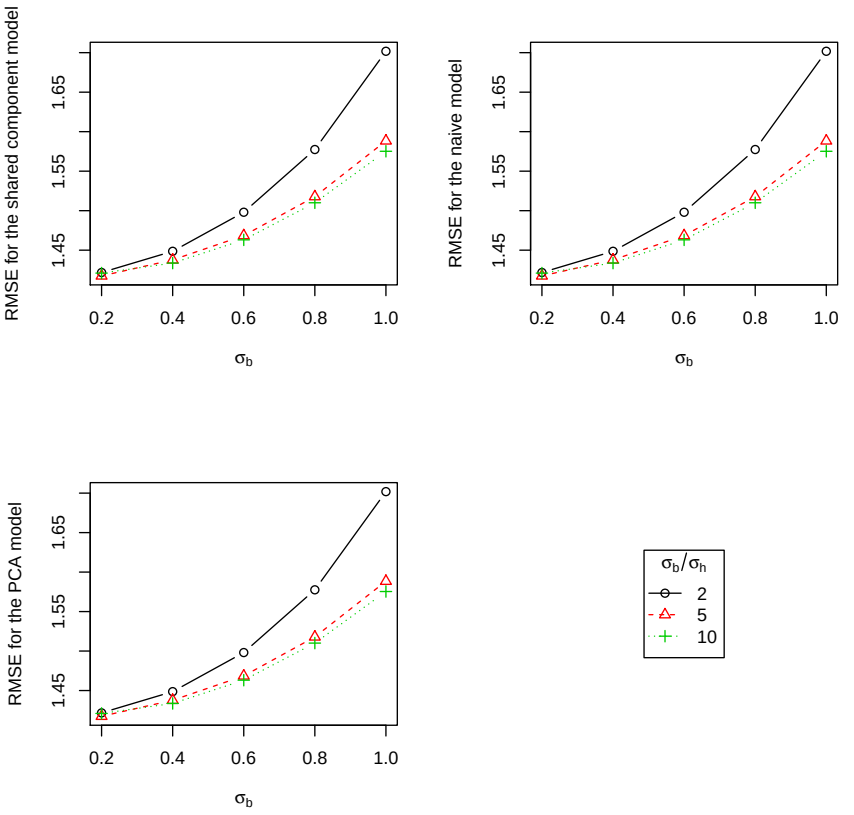


Figure 5.4: The averaged RMSE of the predicted response variable based on the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples.

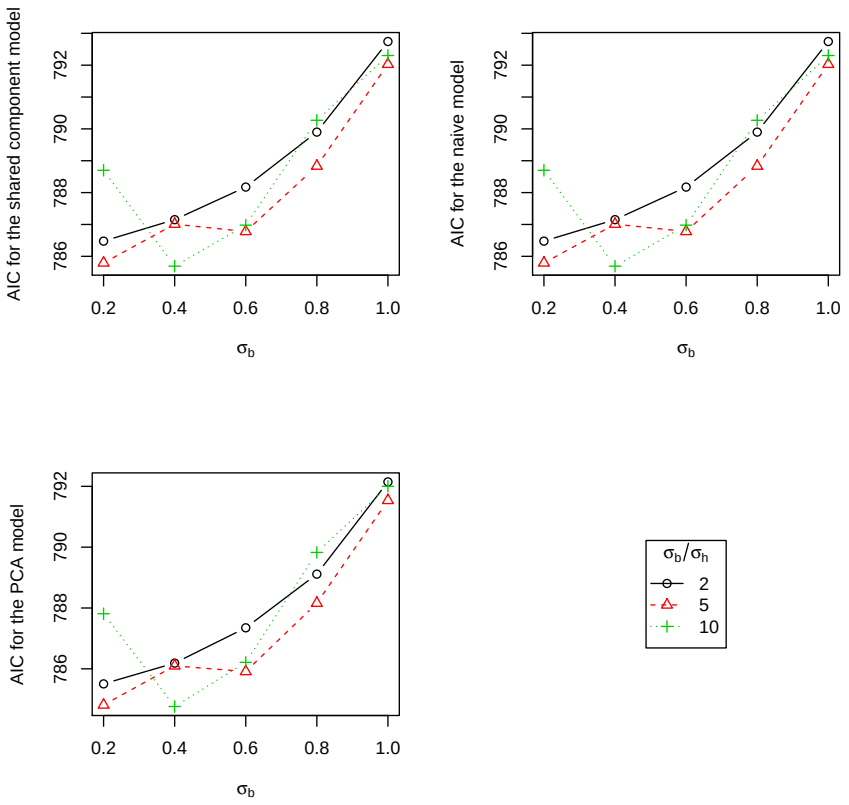


Figure 5.5: The averaged AIC of the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples.

6. Discussion and Future Work

6.1 Discussion

This study proposed a two-stage shared component model with a primary goal of overcoming the issue of multicollinearity among the explanatory variables in a regression analysis. The shared and residual components calculated from the first stage of the proposed model allow the multiple highly correlated explanatory variables to be investigated simultaneously in the same model to predict the outcome. Compared to the single pollutant models which only use partial explanatory information to explain the outcome, the models ignoring the multicollinearity issue by including all the correlated explanatory variables in the model showed slightly better predictive power. However, the existence of multicollinearity can seriously bias the estimation on covariate effect, which will lead to an inaccurate estimation and interpretation of the explanatory variables. The proposed two-stage shared component model is tested on both real and simulated data. Our results showed that the proposed method could correctly estimate the common and residual effects of the explanatory variables.

While the proposed two-stage shared component model is applied, it should be noticed that the values of the latent variables are dependent on the order of the explanatory variables entering the Stage-1 model, especially the first entered variable, which the shared component derive from. Consequently, the coefficient estimation from the two-stage shared component model can vary when different orders are adopted. In our study we prefer to let the covariate with the strongest linear relationship with the others enter the model first, as the resulted shared component can theoretically best represent the correlation among the covariates. The two-stage shared component models with different covariates entered first were also investigated for a comprehensive analysis. From the results of real data analysis shown in Chapter 4 and Appendix A, we can conclude that the findings inferred from the two-stage shared component models with different orders of the explanatory variables entering the model are similar, that the shared component constantly significantly increased the respiratory disease

risk.

Besides the statistical improvement in outcome prediction and coefficient estimation, the proposed two-stage shared component model offers a much easier interpretation on the coefficient estimates from an epidemiological perspective, when compared to other two-stage models like PCA. Unlike the principal components, which are widely recognized hard to interpret, the shared component from our proposed model can stand for the common effect from the multiple covariates, and the residual components show how each covariate performs after the common effect is accounted for. The latent variables help us understand how the outcome is affected by the multiple covariates simultaneously. However, it is challenging to quantify the relationship between the latent variables and the outcome, as each unit of the original explanatory variables differs from the latent variables. The application of our proposed model on the real data finds that the three pollutants $PM_{2.5}$, PM_{10} and NO_2 jointly increase the respiratory disease risk, and NO_2 exhibits an additional stronger negative effect on the respiratory health.

Previous research also used a similar strategy as our proposed two-stage shared component model for modelling the impact of two pollutants, i.e., PM_{10} and NO_2 on the disease outcome [39]. However, few studies have formally evaluated the properties of the two-stage shared component model in comparison to the commonly used methods to tackle the multicollinearity issue among covariates via simulation studies. Further, our proposed model investigated three highly correlated pollutants, which demonstrates that the two-stage shared component model can be used for more than two covariates.

6.2 Limitations and Future Work

As the environmental health data is often collected over geographical regions, it is inevitable there is spatial autocorrelation underlying these data. However, the highly correlated exposures and residual spatial autocorrelation could be confounded [16]. In this case, it is challenging to assess the impact of multicollinearity among the explanatory variables on the disease risk when the residual spatial autocorrelation is present. An improved method to appropriately separate their effect is therefore of interest in future work.

A limitation of this study is that the analysis is based on the yearly environmental and health data, which may not precisely reflect the relationships between the pollutants and the respiratory health. Environmental health data are often measured dynamically on a daily or weekly basis, which could provide more accurate information to assess the relationship between exposures and the health outcome. Extending the proposed two-stage shared component model to analyse time-series environmental health data could be useful and meaningful. Moreover, in environmental studies many other factors are usually considered relating to the concentrations of the air pollutants, such as temperature, humidity [5, 40, 41], traffic [1, 2, 4, 6, 42] and industrial point sources [1, 42], therefore included in the health research to predict disease risk. As we used the retrospective data from a previous study that these relating factors were not collected, the analysis may be limited due to the lack of observations.

Further, in both stages of the proposed method, the linearity of the covariate effects is assumed, which is not violated in this study. However, those assumptions may not be met in other applications. For example, the shared and residual components may be non-linearly associated with the response variable, and in this case, generalized additive model [43] by modelling the effect of the latent components as spline functions would be a more flexible and appropriate approach. The potential interaction effect among pollutants is also a major challenge in evaluating the health impact of pollutant mixture in research [9], which is not modelled in our study that we assume additive effect between the pollutants. A future work to develop models with interactions between the pollutants can be considered.

Bibliography

[1] Le ND, Sun L, Zidek JV. Air pollution. Chronic diseases in Canada. 2010;29(Suppl 2):144–63.

[2] Beelen R, Hoek G, van den Brandt PA, Goldbohm RA, Fischer P, Schouten LJ, et al. Long-term exposure to traffic-related air pollution and lung cancer risk. Epidemiology (Cambridge, Mass). 2008 Sep;19(5):702–710.

[3] Pope CA, Burnett RT, Krewski D, Jerrett M, Shi Y, Calle EE, et al. Cardiovascular mortality and exposure to airborne fine particulate matter and cigarette smoke: shape of the exposure-response relationship. Circulation. 2009 Sep;120(11):941–948.

[4] Zanobetti A, Gold DR, Stone PH, Suh HH, Schwartz J, Coull BA, et al. Reduction in heart rate variability with traffic and air pollution in patients with coronary artery disease. Environmental Health Perspectives. 2010 Mar;118(3):324–330.

[5] Chock DP, Winkler SL, Chen C. A study of the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. Journal of the Air & Waste Management Association (1995). 2000 Aug;50(8):1481–1500.

[6] Crouse DL, Goldberg MS, Ross NA, Chen H, Labrèche F. Postmenopausal breast cancer is associated with exposure to traffic-related air pollution in Montreal, Canada: a case-control study. Environmental Health Perspectives. 2010 Nov;118(11):1578–1583.

[7] Li X, Liu Y, Liu F, Wang Y, Yang X, Yu J, et al. Analysis of short-term and sub-chronic effects of ambient air pollution on preterm birth in central China. Environmental Science and Pollution Research. 2018 Jul;25(19):19028–19039.

[8] Fleischer NL, Merialdi M, van Donkelaar A, Vadillo-Ortega F, Martin RV, Betran AP, et al. Outdoor Air Pollution, Preterm Birth, and Low Birth Weight: Analysis of the World Health Organization Global Survey on Maternal and Perinatal Health. Environmental Health Perspectives. 2014 Apr;122(4):425–430.

[9] Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environmental Health. 2013 Oct;12(1):85.

[10] Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, et al. Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhalation Toxicology. 2010 Jun;22 Suppl 1:1–19.

[11] Graham MH. Confronting Multicollinearity in Ecological Multiple Regression. Ecology. 2003;84(11):2809–2815.

[12] Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology (Sunnyvale, Calif). 2016 Apr;6(2).

[13] Chen GJ. A simple way to deal with multicollinearity. Journal of Applied Statistics. 2012 Sep;39(9):1893–1909.

[14] Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015 Jul;16(3):493–508.

[15] Powell H, Lee D. Modelling spatial variability in concentrations of single pollutants and composite air quality indicators in health effects studies. Journal of the Royal Statistical Society Series A (Statistics in Society). 2014;177(3):607–623.

[16] Lee D, Rushworth A, Sahu SK. A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. Biometrics. 2014 Jun;70(2):419–429.

[17] Kalnins A. Multicollinearity: How common factors cause Type 1 errors in multivariate regression. Strategic Management Journal. 2018 Aug;39(8):2362–2385.

[18] Kutner M. Applied linear statistical models. McGraw-Hill Irwin; 2005.

[19] Myers RH. Classical and modern regression with applications. 2nd ed. Duxbury advanced series in statistics and decision sciences. Boston: PWS-KENT PubCo; 1990.

[20] Lavery MR, Acharya P, Sivo SA, Xu L. Number of predictors and multicollinearity: What are their effects on error and bias in regression? Communications in Statistics - Simulation and Computation. 2019 Jan;48(1):27–38.

[21] Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics. 2010 Jun;13(3):253–267.

[22] O'Brien RM. Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea. Social Science Quarterly. 2017;98(1):360–375.

[23] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901;2(6):559–572.

[24] Neter J. Applied linear regression models. Homewood, Ill.: RDIrwin; 1983.

[25] Matloff NS. Statistical regression and classification: from linear models to machine learning. Texts in statistical science. CRC Press; 2017.

[26] Goldberger AS, Jochems DB. Note on Stepwise Least Squares. Journal of the American Statistical Association. 1961;56(293):105–110.

[27] Jolliffe IT. Principal component analysis. 2nd ed. Springer series in statistics. New York: Springer; 2002.

[28] Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2001;164(1):73–85.

[29] Wang F, Wall MM. Generalized common spatial factor model. Biostatistics (Oxford, England). 2003 Oct;4(4):569–582.

[30] Macnab YC. On Bayesian shared component disease mapping and ecological regression with errors in covariates. Statistics in Medicine. 2010;29(11):1239–1249.

[31] Feng CX, Dean CB. Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. Environmetrics. 2012;23(6):493–508.

[32] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific model development. 2014;7(3):1247–1250.

[33] Arnold TW. Uninformative Parameters and Model Selection Using Akaike's Information Criterion. The Journal of Wildlife Management. 2010;74(6):1175–1178.

[34] Symonds MRE, Moussalli A. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology. 2011 Jan;65(1):13–21.

[35] Altman N, Krzywinski M. Points of Significance: Regression diagnostics. Nature Methods. 2016;13(5):385–386.

[36] Zhou M, Li L, Dunson D, Carin L. Lognormal and Gamma Mixed Negative Binomial Regression. Proc Int Conf Mach Learn. 2012;2012:1343–1350.

[37] Dunn PK, Smyth GK. Randomized Quantile Residuals. Journal of Computational and Graphical Statistics. 1996;5(3):236–244.

[38] Feng C, Li L, Sadeghpour A. A comparison of residual diagnosis tools for diagnosing regression models for count data. BMC Medical Research Methodology. 2020 Jul;20(1):175.

[39] Huang G, Lee D, Scott EM. Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. Statistics in Medicine. 2018;37(7):1134–1148.

[40] Qiu H, Yu ITS, Wang X, Tian L, Tse LA, Wong TW. Cool and dry weather enhances the effects of air pollution on emergency IHD hospital admissions. International journal of cardiology. 2013;168(1):500–505.

[41] Nuvolone D, Balzi D, Chini M, Scala D, Giovannini F, Barchielli A. Short-Term Association Between Ambient Air Pollution and Risk of Hospitalization for Acute Myocardial Infarction: Results of the Cardiovascular Risk and Air Pollution in Tuscany (RISCAT)

Study. American Journal of Epidemiology. 2011 Jul;174(1):63–71. Publisher: Oxford Academic.

[42] Silva R, Adelman Z, Fry M, West J. The Impact of Individual Anthropogenic Emissions Sectors on the Global Burden of Human Mortality due to Ambient Air Pollution. Environmental Health Perspectives. 2016;124(11):1776–1784.

[43] Wood SN. Inference and computation with generalized additive models and their extensions. Test. 2020 Jun;29(2):307–339.

A. Results of the Two-stage Shared Component Model with Different Exposure Variable Entered the Stage 1 Model First

A.1 PM_{10}

X_{i1}, X_{i2}, X_{i3} denote PM_{10}, NO_2, $PM_{2.5}$ respectively in the i^{th} region of the study sample.

Table A.1: The estimated parameters in exponential term and the overall fit of the proposed two-stage model

	Poisson Model		Negative Binomial Model	
	Model 3	Model 4	Model 3	Model 4
e^{β_1}	1.227* (1.134, 1.327)	1.227* (1.134, 1.327)	1.180* (1.028, 1.354)	1.175* (1.024, 1.348)
e^{β_2}	1.446* (1.141, 1.829)	1.446* (1.141, 1.829)	1.550* (1.019, 2.361)	1.354 (0.966, 1.899)
e^{β_3}	0.498 (0.226, 1.098)	0.498 (0.226, 1.098)	0.448 (0.102, 1.957)	-
e^{δ}	2.110* (2.033, 2.191)	2.110* (2.033, 2.191)	2.160* (2.020, 2.309)	2.177* (2.029, 2.324)
AIC	2505.8	2505.8	2258.2	2257.3
-2LL	2495.9	2495.9	2246.2	2247.3

Notes: Model 3: Two-stage shared model; Model 4: Backward selection of the two-stage shared model. The estimated covariate effects are presented as relative risks for one unit increase in each covariates value.

A.2 NO_2

X_{i1}, X_{i2}, X_{i3} denote NO_2, $PM_{2.5}$, PM_{10} respectively in the i^{th} region of the study sample.

Table A.2: The estimated parameters in exponential term and the overall fit of the proposed two-stage model

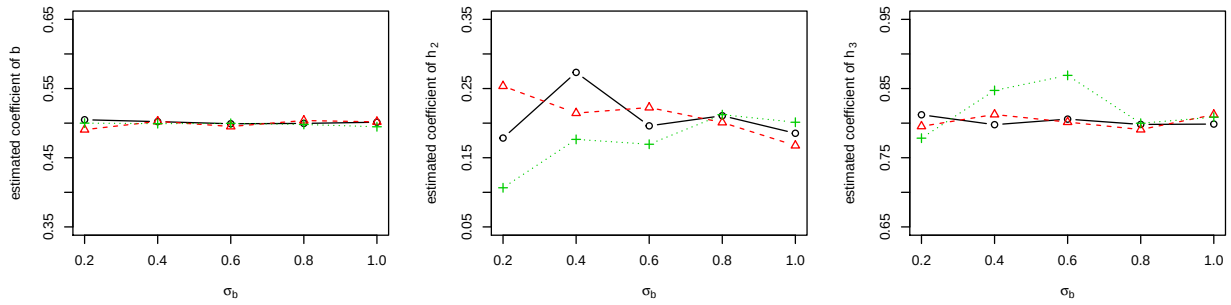
	Poisson Model		Negative Binomial Model	
	Model 3	Model 4	Model 3	Model 4
e^{β_1}	1.136* (1.085, 1.188)	1.134* (1.084, 1.187)	1.114* (1.029, 1.206)	1.113* (1.029, 1.205)
e^{β_2}	0.498 (0.226, 1.098)	0.671* (0.454, 0.994)	0.448 (0.102, 1.957)	0.578 (0.287, 1.160)
e^{β_3}	1.339 (0.683, 2.620)	-	1.275 (0.369, 4.417)	-
e^{δ}	2.110* (2.033, 2.191)	2.116* (2.040, 2.196)	2.160* (2.020, 2.309)	2.165* (2.028, 2.312)
AIC	2505.8	2504.6	2258.2	2256.3
-2LL	2495.9	2496.6	2246.2	2246.3

Notes: Model 3: Two-stage shared model; Model 4: Backward selection of the two-stage shared model. The estimated covariate effects are presented as relative risks for one unit increase in each covariates value.

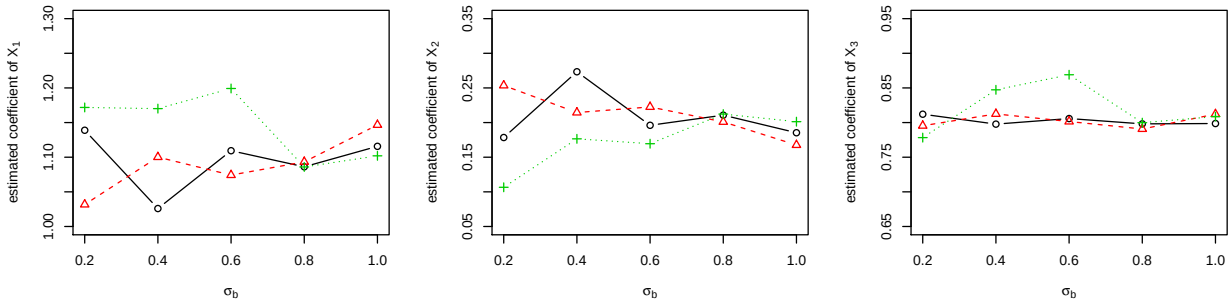
B. Simulation Results with Negatively Correlated Explanatory Variables

Table B.1: The average of the estimated correlation coefficient (r) and variance inflation factor (VIF) for the explanatory variables from the 1000 simulated samples.

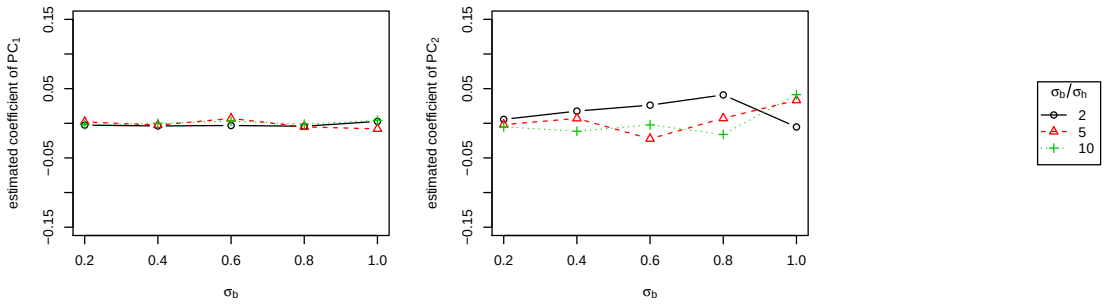
σ_b/σ_{h3}	r_{12}	r_{13}	r_{23}	VIF_1	VIF_2	VIF_3
2	0.995	-0.894	-0.890	107	102	5
5	0.995	-0.981	-0.976	128	102	26
10	0.995	-0.995	-0.990	204	102	102



(a) The average of the estimated regression coefficients of the naive model, i.e., the Poisson regression including all the explanatory variables in the model.

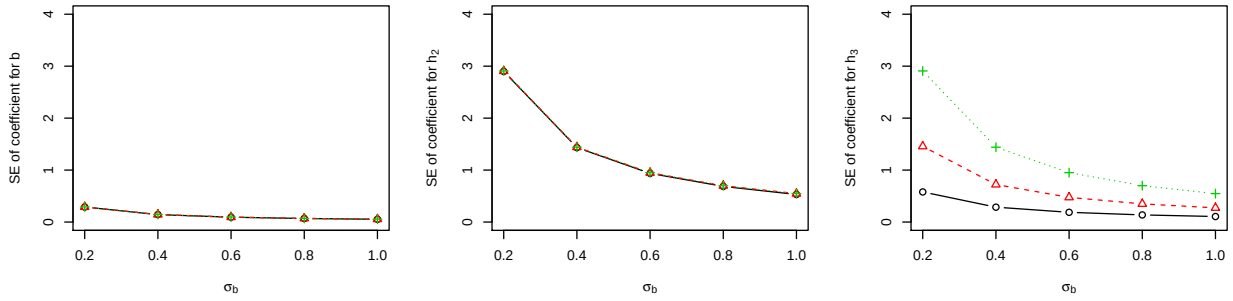


(b) The average of the estimated regression coefficients of the naive model, i.e., the Poisson regression including all the explanatory variables in the model.

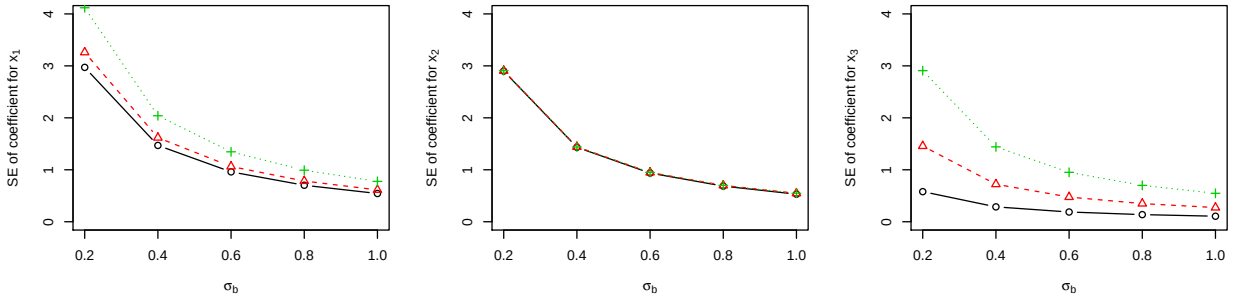


(c) The average of the estimated regression coefficients of the two-stage PCA model, i.e., the Poisson regression including the principal components of the explanatory variables.

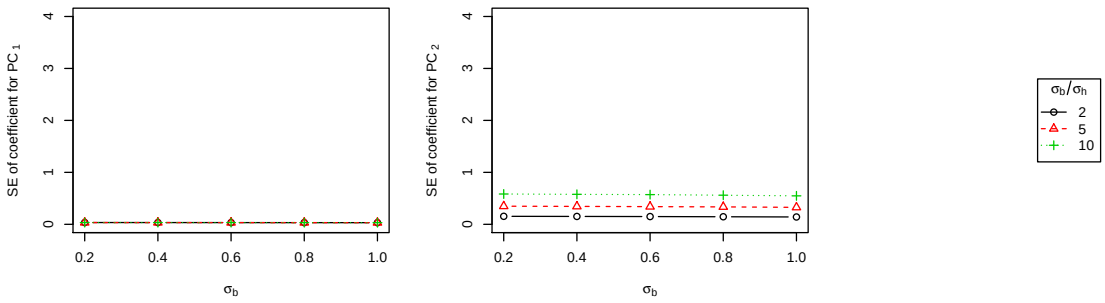
Figure B.1: The average of the estimated negative regression coefficients of the proposed two-stage shared component model (top panel), the naive model (middle panel) and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.



(a) The average of the standard errors of the estimated regression coefficients for the two-stage shared component model.

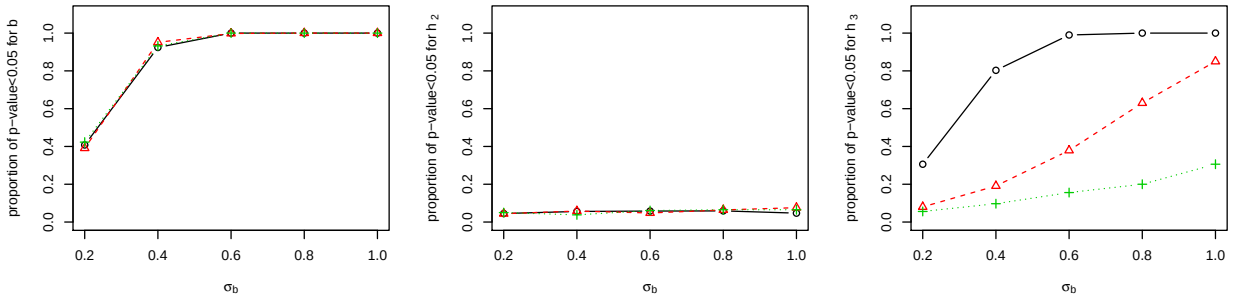


(b) The average of the standard errors of the estimated regression coefficients for the naive model.

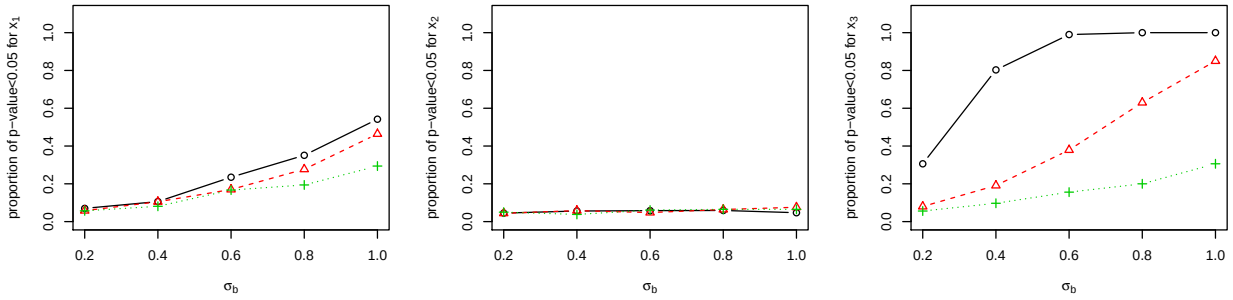


(c) The average of the standard errors of the estimated regression coefficients for the two-stage PCA model.

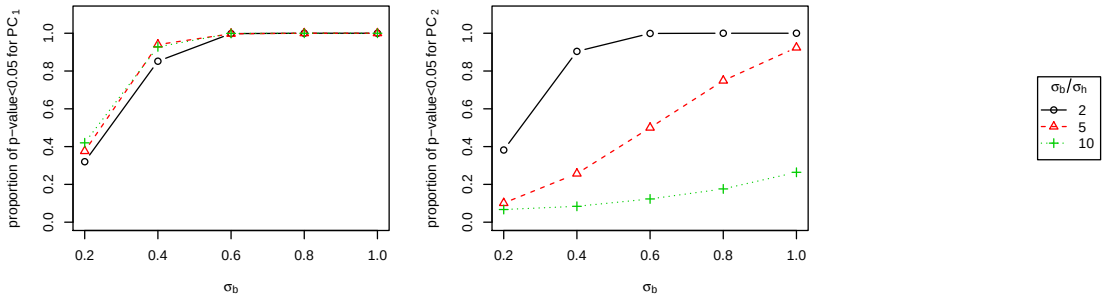
Figure B.2: The average of the standard errors of the regression coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.



(a) The probability of statistically significant coefficients for the two-stage shared component model.



(b) The probability of statistically significant coefficients for the naive model.



(c) The probability of statistically significant coefficients for the two-stage PCA model

Figure B.3: The probability of statistically significant coefficients for the proposed two-stage shared component model (top panel), the naive model (middle panel), and the two-stage PCA model (bottom panel) based on the 1000 simulated samples.

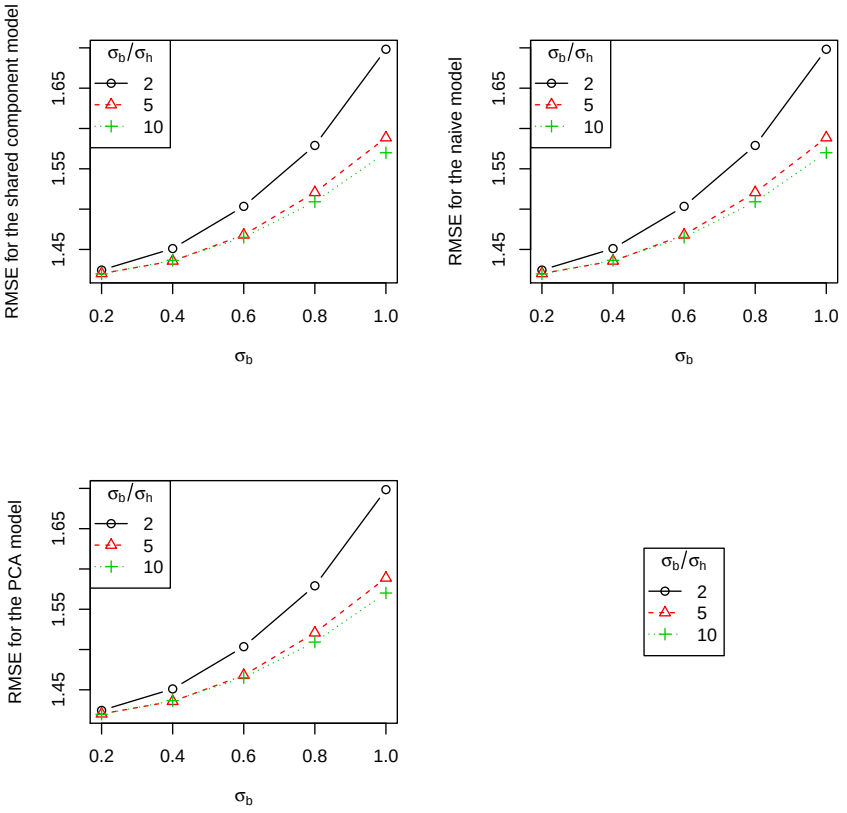


Figure B.4: The averaged RMSE of the predicted response variable based on the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples.

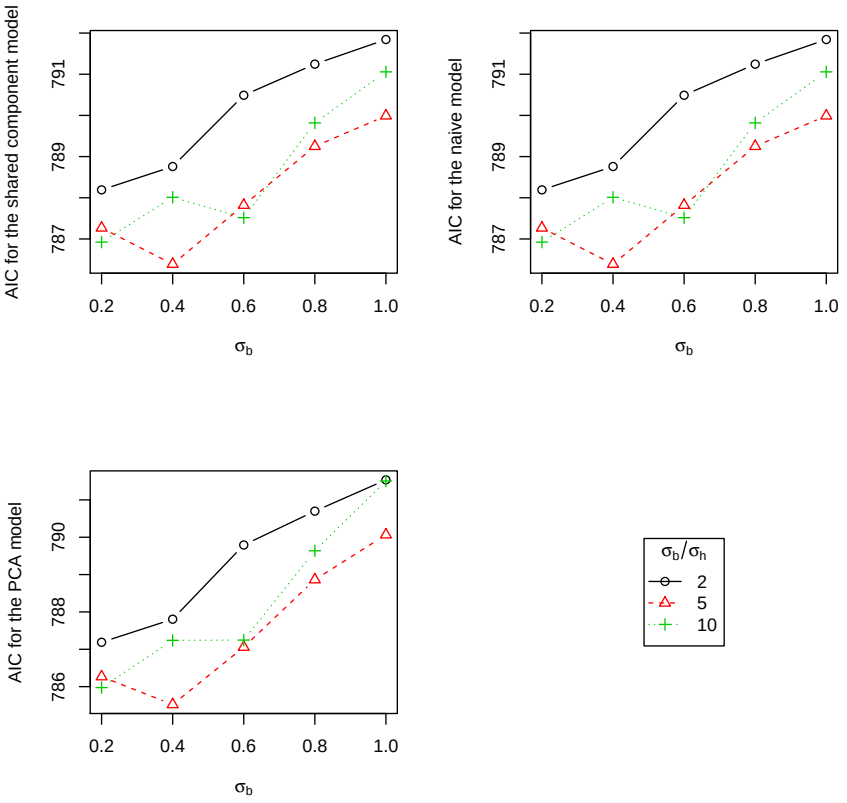


Figure B.5: The averaged AIC of the proposed two-stage shared component model, the naive model, and the two-stage PCA models from 1000 simulated samples.

C. Research Ethics Approval

UNIVERSITY OF SASKATCHEWAN

Research Services and Ethics Office
University of Saskatchewan
Room 223 Thorvaldson Building
110 Science Place
Saskatoon SK Canada S7N 4J8

September 12th, 2019

Dr. Cindy Feng
School of Public Health
University of Saskatchewan
Saskatoon Sk

U of S BIO: 1434

Dear Dr. Feng,

Thank you for your application entitled, "Bayesian Latent Variable Models for Modeling the Associations of Multiple Exposures in Relationship to Adverse Health Outcomes" (Bio ID 1434). In the opinion of the Research Ethics Board (REB) this submission is exempt from the requirement of Research Ethics Board (REB) review and approval based on article 2.5 of the Tri-Council Policy Statement (TCPS2). Article 2.5 specifies *"quality assurance and quality improvement studies, program evaluation activities, and performance reviews, or testing with normal educational requirements when used exclusively for assessment, management or improvement purposes do not constitute research for the purposes of this Policy, and do not fall within the scope of REB review."*

Although this project is exempt of the requirement for research ethics review, it should be conducted in an ethical manner in accordance with the information that you submitted to the REB and in keeping with the Saskatchewan Health Information Protection Act (HIPA). Any deviation from the original methodology should be brought to the attention of the Biomedical Research Ethics Board for further review.

Digitally Approved by Gordon McKay, Ph.D.,
Chair, Biomedical Research Ethics Board
University of Saskatchewan